Transkriptomidatan soveltuvuus populaatiogeneettisiin analyyseihin: tapaustutkimuksena kahden suomalaisen kiiltomatopopulaation (*Lampyris noctiluca*) analyysi

Santtu Urpilainen

Pro gradu -tutkielma

Oulun yliopisto

Biologian tutkinto-ohjelma

30.11.2020

Sisällys

Tiivistelmä

Populaatiogenomiikka on ollut kasvava tutkimuksen ala viimeiset 20 vuotta ja nykyään use-alle ei-malliorganismille on sekvensoitu genomisekvenssi. Varsinkin transkriptomiikka on noussut esille tehokkaana menetelmänä ei-malliorganismien sekvensoinnissa, koska se mahdollistaa geneettisen muuntelun tutkimuksen ilman koko genomin sekvensoinnin aiheuttamia kuluja ja aikaa vieviä analyysejä. Bioinformatiikan osaaminen on noussut tärkeäksi osaksi populaatiogenetiikan tutkimuksia ja varsinkin aineiston suodattaminen on tärkeää. RNA:n sekvensoinnin suurena ongelmana on saatujen sekvenssien. Sen lisäksi transkriptomit ovat monimutkaisia ja niissä on suurena ongelmana piilotettu paralogia, jolloin duplikaatin menetyksen jälkeen geeni tunnistetaan virheellisesti yhdenkopion ortologiksi.

Kiiltomato (*Lampyris noctiluca*) on yöelämään sopeutunut kovakuoriainen, jonka naaraat tuottavat luminesenssia loistetta takaruumiinsa valoa tuottavalla elimellä. Kiiltomadon esiintyvyys on laaja aina Euroopan länsirannikosta Kiinaan asti, mutta sen elinympäristö on rajoittunut ihmisen toiminnan seurauksena. Lajin sukupuolidimorfismi on ollut sen luminesenssin ohella yleinen tutkimuksenaihe ja aikaisempaa populaatiogeneettistä tutkimusta ei ole tehty. Tämän seurauksena kiiltomadon genetiikasta ei ole paljoakaan ymmärrystä. Kiiltomatonaaraiden tiedetään liikkuvan vähän aikuisina ja vain koiraat kykenevät lentämään, mikä tekee sen populaatiorakenteen tutkimisen mielenkiintoiseksi.

Tämän pro gradu -tutkielman päämääränä oli selvittää miten paralogien suodatus vaikuttaa transkriptomin soveltuvuuteen populaatiogeneettisiin analyyseihin tutkien Viljakainen ym. 2020 tutkimuksessa kerättyjen Hangon ja Konneveden *L. noctiluca* -populaatioiden välistä geneettistä muuntelua. Tutkielmassa tehtiin sekvenssilukemien linjaus koostettuihin transkriptomeihin ja suodatus ennen populaatiogeneettisiä analyysejä. Analyysit tehtiin yleisellä tasolla suodatetulle aineistolle sekä McKinneyn-menetelmällä suodatetulle aineistolle.

Ennen suodattamista aineiston populaatiogeneettisistä analyyseistä havaittiin korkeaa heterotsygotiaa ja vääristymää alleelifrekvensseissä. McKinneyn suodattamisen seurauksena havaitut korkean heterotsygotian vääristymät poistuivat aineistosta. Syynä tähän on todennäköisesti aineiston piilotettua paralogiaa ja siitä johtuvaa paralogien ylijäämää. *L. noctiluca* populaatioiden välillä oli vähän muuntelua ja niissä ei havaittu sukusiittoisuutta. Populaatioiden välillä, kun tarkastelu rajattiin koiraisiin kummassakaan populaatiossa ei esiintynyt lainkaan muuntelua. Syytä populaatioiden vähäiselle geneettiselle muuntelulle ei tiedetä.

1. Johdanto

Populaatiogenomiikka perustuu lajin sisäisen molekyylimuuntelun analyysiin, joka on viimeisen vuosikymmenen aikana ollut nouseva tutkimuksen osa-alue (Gayral ym. 2013). RNA-tutkimus mullistui vuonna 1990 kun suuressa mittakaavassa tehtävän RNA-sekvenssien tutkimuksen mahdollistavat suurenkapasiteetin menetelmät yleistyivät. Viimeisen 20 vuoden aikana uuden sukupolven sekvensoinnilla on saatu kerättyä paljon uutta tietoa DNA:sta ja RNA:sta (Hrdlickova ym. 2017). Tämä johtuu sekvensoinnin kustannuksien laskusta ja siitä, että genomisen aineiston tuottaminen on nykyään mahdollista uuden sukupolven sekvensoinnin tekniikoilla. Näitä käytetään tuottamaan suuria määriä geneettistä informaatiota genomista tai transkriptomista populaatiogeneettisiin tutkimuksiin (Haas ym 2013; Ellegren 2014; Wright ym. 2019). Varhaisimmat populaatiogenomiikan tutkimukset kohdistuivat pääasiassa pieneen määrään malliorganismeja, kuten esimerkiksi hiivaan (*Saccharomyces cerevisiae*) ja lituruohoon (*Arabidobsis thaliana*) sekä eläintutkimuksissa ihmiseen ja banaanikärpäseen (*Drosophila melanogaster*), joista on olemassa referenssigenomit (Gayral ym. 2013). Nykyään genomisekvenssit eivät enää ole rajoittuneita vain malliorganismeihin, mikä mahdollistaa usean globaalille ekologialle ja evoluutiobiologialle tärkeän lajin tutkimuksen ja mahdollistaa populaatiogenomiset projektit kuten, Earth Biogenome Project, joka pyrkii kartoittamaan kaiken eukaryoottisen elämän. (Hudson, 2008; Haas ym. 2013; Wright ym. 2019). Uudemmissa populaatiogenomiikan tutkimuksissa on tutkittu muun muassa sinivalaita (Attard ym. 2018), verifasaaneja (Zhou ym. 2020) ja susia, joista on koostettu referenssigenomi (Gopalakrishnan ym. 2017).

Transkriptomi-tutkimusta on tehty lukuisille eliöille muun muassa useille kasveille (Gautier ym. 2017) sekä kiiltomadoille, kuten yhdelle *Rhagophthalmidae* -lahkon hyönteiselle (*Rhagophthalmus sp.*) ja kolmelle tulikärpäselle (*Asymmetricata circumdata*, *Aquatica ficta*, ja *Pyrocoelia pectoralis*) fylogeneettisessä tutkimuksessa, missä osoitettiin *Rhagophthalmidae* -lahkon eroavan *Lampyridae* -lahkosta (Wang ym. 2017).

Uuden sukupolven sekvensointitekniikat mahdollistavat ei-malliorganismien populaatiogenomiikan (Gayral ym. 2013) ja varsinkin RNA-sekvensointimenetelmää (RNAseq) käytetään lajien transkriptomien karakterisointiin. Sen suurena etuna muihin uuden sukupolven sekvensointitekniikoihin on se, että se mahdollistaa populaatioiden välisen geenieksprression analysoinnin sekä geneettisen muuntelun tutkimisen kustannustehokkaammin kuin koko genomin sekvensoinnilla. RNA-sekvensoinnin sekvenssidata koostuu tyypillisesti satunnaisista lyhyistä, enintään noin parista sadasta emäsparista koostuvista lukemapareista.

Referenssigenomin tai transkriptomi-informaation puuttuessa transkriptit koostetaan lukemapareista *de novo*, mistä tunnistetaan pistemutaatioiden aiheuttamia polymorfisia paikkoja eli yhden nukleotidin polymorfisia paikkoja (SNP) (Wolf, 2013; Cutter, 2019). Ongelmana lyhyiden lukemien kanssa on kuitenkin se, että niillä saadaan koostettua vain pieniä osia monimutkaisista eukaryoottien genomeista/transkriptomeista riippumatta niiden sekvensoinnin matalasta hintakynnyksestä ja sen nopeudesta. Tehokkain lyhyen lukeman menetelmien käyttö on vertailu hyvälaatuiseen genomidataan, mistä on hyötyä pääasiassa vain täysin sekvensoiduille lajeille ja niiden sukulaisille. Kuitenkin transkriptomin sekvensoinnilla on mahdollista tehdä tutkimusta ei-malliorganismeilla ilman olemassa olevia genomisekvenssejä (Hudson, 2008).

1.1. Kiiltomato (Lampyris noctiluca)

Lampyris noctiluca on yöelämään sopeutunut kovakuoriainen, jonka naaraat kykenevät luomaan vihertävää valoa takaruumiin valoa tuottavalla elimellä ja toukat kylkiensä pisteillä. Kiiltomadon bioluminesenssi toimii jo toukkavaiheessa aposemaattisena varoituksena saalistajille, joko niiden myrkyllisyydestä tai pahasta mausta, mistä esimerkiksi rupikonnat oppivat olemaan koskematta hohtaviin saaliisiin. Naaraalla bioluminesenssi toimii houkuttimena lentokykyisille koiraille (Viljakainen ym. 2020; De Cock & Matthysen, 2003; Tyler 2002). Kiiltomato on laajasti levinnyt pohjoisen pallonpuoliskon halki aina Espanjasta Kiinaan ja Kreikasta Suomeen asti. Kuitenkin naaraiden lentokyvyttömyydestä ja toukkien kuoriutumisaikaeroista johtuen kiiltomatopopulaatiot ovat eristyksissä toisistaan (Tyler 2002).

Bioluminesenssi on useissa eri organismeissa esiintyvä huomiota herättävä prosessi, joka on havaittavissa useilla kalalajeilla, maa- ja meriselkärangattomilla, sienillä, bakteereilla ja alkueläimillä. Luonnossa bioluminesenssillä on useita käyttötarkoituksia: se toimii suojautumiskeinona, houkuttimena, puolustuksena, varoituksena, kommunikaationa, matkimisena ja valontuottajana (Wang ym. 2017). Kovakuoriaisissa bioluminesenssi perustuu lusiferaasi entsyymiin ja sen substraattiin, lusiferiiniin, joka on tähän asti karakterisoitu kuudesta eri *Lampyridae* -heimon kovakuoriaisesta (Day ym. 2006). *L. noctilucan* on havaittu loistavan heikosti muullakin kuin takaruumiinsa valoa tuottavalla elimellä. Toukka- ja aikuisvaiheessa se loistaa myös jaokkeissa ja läpinäkyvissä kohdissa kehon reunoilla sekä koiraalla visiirissä. On mahdollista, että tämä luminesenssivuoto johtuu alkeellisesta valoelimen morfologiasta, joka ei enää ole kiiltomadon hallinnassa. Kuitenkaan saalistajat ja toiset kiiltomadot eivät vaikuta näkevän tätä luminesenssivuotoa, joten se ei toimi aposemaattisena tai seksuaalisena signaalina (Tisi ym. 2014).

Bioluminesenssin omaavista hyönteisistä käytetään tyypillisesti peittotermiä kiiltomato, mutta tähän määritelmään lasketaan *Collembola* (hyppyhäntäiset), *Diptera* (kärpäset), *Coleoptera* (kovakuoriaiset) -lahkojen hyönteiset. Kovakuoriaisissa bioluminesenssin omaavia heimoja ovat tulikärpäset (*Lampyridae*), rautatiemadot (*Phengodidae*), sepät (*Elateriadae*) ja *Rhagophthalmidae*-heimon kovakuoriaiset (Wang ym. 2017). Kovakuoriaisilla, kuten sylkikuoriaisilla (*Cartharidae*) ja tulikärpäsillä (*Lampyridae*) aposemaattinen vaste esiintyy varoituksena saalistajille pahasta mausta ja hajusta. Näiden lajien aposemaattinen vaste on mahdollisesti kehittynyt erikseen tai siirtynyt yhteisen esi-isän aposemaattisen värityksen mukana. Tämän yhteisen esi-isän kautta tulikärpäsille kehittyi myöhemmin aposemaattiseen vasteeseen pohjautuva bioluminenssi, joka myöhemmin tuli osaksi niiden parinvalintaa (Oba ym. 2019; Sagegami-Oba ym. 2007).

1.1.1. Kiiltomadon elinkaari

Kiiltomadon munilla kestää 27–45 päivään kuoriutua riippuen ympäristön lämpötilasta. Luminesenssia voi esiintyä muutamana ensimmäisenä päivänä munimisesta, jolloin syntyy heikkoa kellertävää valoa. Kiiltomadon toukkien on havaittu kuoriutuvan yöllä, ja toukat ovat kuoriutuessaan noin viiden millimetrin pituisia ja niiden kuori on väriltään savun harmaa, mutta muutaman tunnin jälkeen kuoren kovettuessa se tummentuu, kunnes se on suureksi osaksi täysin musta lukuun ottamatta valkeita läiskiä segmenttien reunoilla (Tyler 2002).

Toukkavaiheessa koiraalla ja naaraalla ei ole mitään näkyvää eroa. Kuoriutumista seuraavan vuoden aikana toukkien koko kasvaa dramaattisesti. Kun ne ovat lähes täysikasvuisia, ne painavat noin 100 kertaa enemmän kuin syntyessään. Kuitenkaan toukan ulkonäkö ei muutu, vaan suuret toukat ovat suurennettuja versioita pienistä (Tyler 2002).

Kiiltomadot eivät syö aikuisvaiheessa, mutta toukkina niiden ruokavalio koostuu suurimmaksi osaksi etanoista. Kuitenkin kiiltomadon toukat pystyvät syömään myös muuta ruokaa, kuten muita hyönteisiä. Kiiltomadon toukka lähestyy saalistaan hiipimällä ja tarkkailee tätä ennen hyökkäystä. Toukan aseena toimii sen suuosien kautta saaliiseen siirtyvä myrkky, joka halvaannuttaa ja sulattaa kohteensa. Suuremmat kohteet tarvitsevat enemmän puremia, jotta myrkky vaikuttaa (Tyler 2002).

Toukkavaihe kestää kaksi vuotta ja jossain tapauksessa kolme vuotta, mikä saattaa johtua siitä, että toukat talvehtivat kahdesti. Koiraan ja naaraan kotelovaiheen pituudessa on eroa noin 5 päivää ja aikuistuneiden yksilöt ilmestyvät kesä-heinäkuussa (Tyler 2002). On havaittu, että ravinnon tarpeen seurauksena naarailla toukkavaihe voi kestää kolmekin vuotta

(Horne & Horne 2017). Yleisesti kiiltomatojen elinkaari kestää 24 kuukautta, mikä jakautuu kolmeen kalenterivuoteen ja kahteen horrostilaan talvella. Ensimmäisen horroksen jälkeen toukat aloittavat ravinnon hankinnan ja yleensä toukkien kasvu on nopeinta ensimmäisinä viitenä kuukautena (Tyler 2002; Horne 2011). Loistekauden loppuvaiheessa koteloitumiseen sopivien toukkien on huomattu horrostavan yhden ylimääräisen talven ja koteloituvan seuraavana vuonna. On mahdollista, että nämä kynnysvaiheen toukat voivat koteloitua välittömästi sopivassa ympäristössä tai ilmastossa (Hickmott & Tyler, 2011).

Huhtikuun ja kesäkuun aikana kiiltomadon kahden ja kolmen kalenterivuoden vanhojen toukkien on havaittu lähtevän vaeltamaan auringon valossa ja ne kulkevat noin 24 tuntia päivässä. Tähän toimintaan ei tiedetä syytä, mutta tämän vaelluksen syyksi on oletettu se, että näin toukat levittäytyvät uusille elinympäristöille. Naaraat näet liikkuvat vain vähän aikuisuudessa, joten levittäytyminen on mahdollista vain toukkavaiheessa (Tyler 2002).

1.1.2. Kiiltomadon sukupuolidimorfismi

Toukkavaiheen jälkeen kiiltomato muuntautuu sukupuolen mukaan joko lentokykyiseksi koiraaksi tai valoa tuottavaksi naaraaksi. Aikuisvaiheessa olevilla kiiltomadoilta puuttuu suuosat ja lyhyen elinikänsä aikana ne eivät ruokaile. Naaras muistuttaa pintapuoleisesti toukkaa, mutta siltä puuttuvat valkeat täplät jaokkeiden laidoista. Naaraiden koko vaihtelee suuresti välillä 38–360 milligramman välillä. Suuremmat naaraat munivat enemmän johtuen mahdollisesti niiden toukkavaiheessa horrostamasta ylimääräisestä vuodesta. Kiiltomadot ovat saaneet nimensä juuri naaraiden kyvystä tuottaa bioluminenssia valoa. Naaraat tyypillisesti loistavat yhden yön tai enemmän, jos ne eivät pariudu ensimmäisenä loistokertana. Joidenkin naaraiden on nähty loistavan myös koiraiden kadottua, mikä mahdollisesti johtuu epäonnekkuudesta ravinnon saannin suhteen ennen koteloitumista (Tyler 2002; Horne ym. 2017; Horne 2011).

Naaraat ehtivät keskimäärin loistaa kaksi tuntia ennen pariutumista (Hickmott & Tyler, 2011). Naaraiden lisääntymismahdollisuudet paranevat etäisyyden kasvaessa kilpailijoihinsa, minkä vuoksi naaraat ovat levittäytyneet laajalle luonnossa. Naaraat pyrkivät pysymään vähintään noin 10.5 cm – 100 cm etäisyydellä muista loistavista naaraista. Naaraiden ei havaittu loisivan toisten naaraiden loistolla, vaan ne suosivat etäisyyttä (Borshakovski ym. 2019). Naaraat, jotka pysyvät erossa toisista naaraista, lisääntyvät nopeammin ja voivat valita kumppaninsa. Vastaavasti toisiaan lähellä olevat naaraat saavat kumppanin helpommin, kun niiden naapureina on vähemmän puoleensavetäviä yksilöitä (Lehtonen & Kaitala, 2020). Naaraat eivät liiku loistopaikaltaan 30 cm kauemmaksi ja ne suosivat paikkoja, joissa ei ole katosta.

Kiiltomadon valonäytöskausi kestää pisimpään 36 yötä ja alkaa 9 yötä muodonmuutoksesta. Tämän kauden pituus kasvaa naaraiden määrän noustessa (Hickmott & Tyler, 2011).

Sää- ja ympäristöolosuhteet vaikuttavat sekä valonäytöskauden alkuun että koteloitumisen ajankohtaan. Valoisuus ja siten päivänpituus vaikuttavat pääasiassa koteloitumiseen, mutta myös siihen, kuinka kauan naaraat ehtivät loistaa (Hickmott & Tyler, 2011). Nykyään kaupungistumisen seurauksena keinovalo on yleistynyt kiiltomatojen elinympäristöissä. Muutokset elinympäristöissä rajoittavat kiiltomatojen valonäytöskautta ja viimeaikainen tutkimus on osoittanut, että naaraat eivät vaikuta valitsevan loistopaikkaansa keinovalon pohjalta, vaan ne piiloutuvat valoisuuden ajaksi, vaikka valoton alue olisi lyhyen matkan päässä. Niiden toiminta johtuu siitä, että ne valitsevat koteloitumis- ja loistepaikat päivällä vaelluksellaan toukkina. Koiraat vastaavasti eivät havaitse naaraita keinovalon valaisemilta alueilta vaan välttelevät näitä (Elgert ym. 2020; Ineichen & Rüttimann, 2012). Lämpötila vaikuttaa kuivina kausina toukkien ruuan saantiin vähentämällä etanoiden määrää. Vastaavasti kostea ilmasto ja tuulinen sää vaikeuttavat koiraiden kykyä lentää ja siten naaraat loistavat kauemmin (Hickmott & Tyler, 2011).

Koiras eroaa ulkomuodoltaan naaraasta merkittävästi, minkä seurauksena *L. noctiluca* -koirasta ja naarasta pidettiin aikaisemmin eri lajeina. Toisin kuin naaraat koiraat kykenevät lentämään ja ne myös näkevät paremmin kuin naaraat ja toukat. Niiden fysiologiset erot johtuvat niiden aktiivisuudesta lisääntymisessä ja parinvalinnassa. Koiraalla on toukkaan ja naaraaseen verrattuna erittäin hyvä näköaisti. Naaraalla on 300 silmäkeilaa, niin vastaavasti koiraalla on 2 000 keilaa (Tyler 2002). Vaikka koiras kykenee havaitsemaan naaraan taivaalta, sillä on kuitenkin vaikeuksia nähdä kirkkauseroa lyhyillä etäisyyksillä (Borshakovski ym. 2019).

Koiras etsii naaraita noin metrin korkeudella maasta tarkkaillen ruohikosta loistavia naaraita, mutta tutkimuksissa on huomattu, että valon värillä ei niinkään ole merkitystä koiraalle vaan sen kirkkaudella (Tyler 2002). Kiiltomatonaaraiden koolla on vaikutusta niiden fertiliteettiin, mutta koiraiden on vaikeaa havaita tätä, joten naaraat osoittavat sen kirkkaudellaan (Hopkins ym. 2015). Havaitessaan naaraan koiras laskeutuu yllättävän tarkasti ja kulkee loppumatkan jalan. Koiraiden on huomattu taistelevan lisääntyäkseen, mutta naaraat antavat useamman koiraan paritella (Tyler 2002).

1.2. Populaatiogeneettiset menetelmät

1.2.1. Bioinformatiikka

Nykyään bioinformatiikan taidot ja sen ymmärtäminen ovat tärkeitä suurten DNA-sekvenssidatojen analysoinnissa. Bioinformatiikka on ollut kasvava ala 1990-luvulta lähtien ja sen tärkeys ja skaala ovat kasvaneet Human Genome -projektin ja DNA-sekvensoinnin yleistymisen myötä. Viime aikoina toisen ja kolmannen polven sekvensointimenetelmien nousun myötä datan määrä on kasvanut eksponentiaalisesti tutkimuksissa, mikä nostaa bioinformaattisten menetelmien osaamisen tarvetta (Luikart ym. 2018).

Populaatiogeneettisissä tutkimuksissa on kriittisiä vaiheita, jotka suositellaan tekemään. Ensiksi suunnitellaan sekvensoinnin strategia etukäteen. Tämän jälkeen sekvenssidata tuotetaan kerätyistä näytteistä ja suodatetaan paralogien ja laadun suhteen, joilla on vaikutusta tuloksiin. Sitten suodatetut sekvenssidatan lukemat linjataan transkriptomiin tai genomiin, joka tyypillisesti on lajin referenssigenomi, mutta sen puuttuessa on suositeltavaa, että käytetään läheisen lajin referenssigenomia tai *de novo* transkriptomia. Sitten varianttialleelit, kuten SNPt sekä pienet insertiot ja deleetiot, tunnistetaan ja genotyypitetään, jotta saadaan selville niiden genotyyppien todennäköisyydet ja lukusyvyys, eli kuinka useat lukemista linjautuvat referenssiin. Lopuksi on hyvä suodattaa lokukset, jotka poikkeavat biologisista odotuksista muun muassa korkean havaitun heterotsygotian tai poikkeavan alleelimäärän suhteen ennen jatkoanalyysejä (Góngora-Castillo & Buell, 2013; Ellegren 2014; Benestan ym. 2016; Luikart ym. 2018).

1.2.2. Transkriptomiikka

Transkriptomin tutkimuksessa tutkitaan kaikkia näytteenottohetkellä genomista tuotettuja transkriptejä, jotka muodostavat transkriptomin. Populaatiotranskriptomiikassa näiden transkriptien avulla voidaan tutkia populaatioiden sisäisen ja populaatioiden välisen geenien ekspression vaihtelua. Populaatiotranskriptomiikan tutkimukseen käytetään mikrosiruanalyysiä ja RNA-sekvensointia (RNAseq). Hybridisaatioon perustuvat mikrosiruanalyysit ovat korvautumassa RNAseq:llä, koska uuden sukupolven sekvensointi on edullista ja koska RNAseq:llä voidaan tunnistaa ja kvantifioida isoformien ja tuntemattomien transkriptien ekspressiota. RNAseq:llä voidaan koostaa transkriptomi ilman referenssigenomia tai tiettyä sekvenssiä, mikä mahdollistaa transkriptomin rekonstruoinnin *de novo* sekä yhden nukleotidin polymorfisten paikkojen tunnistamisen ja genotyypityksen (Costa-Silva ym. 2017; Luikart ym. 2018).

Transkriptomin datan tutkimuksessa RNAseq:n myötä saaduissa referenssittömissä lyhyissä sekvensseissä on omat ongelmansa. Nämä ongelmat pääasiassa jakautuvat lyhyydestä johtuviin virheisiin ja korkean peittävyyden käyttämiseen satunnaisten sekvensointivirheiden kompensoinnissa, joista aiheutuu kuluja. Tämän lisäksi suuri osa transkriptien koostamisen algoritmeista ei ota laatua huomioon, jolloin sekvenssien lyhyys rajaa tarkkuutta. Analyysiin käytettävät algoritmit ovat myös RAM intensiivisiä. Lopuksi transkriptomit ovat monimutkaisia ja niiden laadun määrittämisen menetelmät ovat rajattuja (Góngora-Castillo & Buell, 2013).

Transkriptomi analyyseissä suurena ongelmana on piilotettu paralogia. Se näkyy polymorfisissa paikoissa siten, että yksilöissä esiintyy tavallista korkeampaa heterotsygotiaa ja yksilöt jakavat yhteisen korkeasti ekspressoidun alleelin (Gayral ym. 2013). Piilotettu paralogia pohjautuu molekyylifylogeneettiseen tutkimukseen, missä geeniparit tunnistetaan ensin ortologisiksi, mutta myöhemmin tunnistetaan paralogisiksi. Nämä piilotetut paralogit sisältävät yhden kopion per laji johtuen duplikaattien menetyksestä, mikä johtaa niiden virheelliseen tunnistukseen yhdenkopion ortologeiksi. (Kuraku 2010; Smith & Hahn, 2020). Esimerkkinä piilotetusta paralogiasta on seeprakalan homeobox-geeni, joka nykyään tunnetaan Emx3-geeninä. Se alun perin tunnistettiin vuonna 1995, kun Emx1 ja Emx2 geenit kloonattiin ja niiden tunnistettiin vaikuttavan aivoissa. Kuitenkin silloisen Emx1-geenin paralogia tetrapoda-yläluokan Emx1 geenin kanssa tunnistettiin jo vuonna 1997 ja vasta vuonna 2002 se tunnistettiin Emx-geeniperheen Emx3-geeniksi ja seeprakalan todellinen Emx1-geenin ortologi tunnistettiin. Emx3-geenin myöhäisen tunnistuksen syynä oli Emx3:n ortologien puuttuminen ihmiseltä, hiireltä ja kanalta. Muita piilotettuja paralogioita ovat ParaHox-geeniryppään Cdx-geeni ja eläimen kehitykseen liittyvät Bmp2 ja Bmp4 geenit, joiden ortologi on kärpäsen dpp-geeni (Kuraku 2010).

1.2.3. Hardyn-Weinbergin laki

Hardyn-Weinbergin tasapaino toimii pohjustuksena valinnan tutkimuksiin ja se perustuu siihen, että satunnaisesti parittelevan populaation genotyypit jakautuvat tasapainotilaan hetero- ja homotsygoottifrekvenssiensä suhteen, jos ei oteta huomioon niiden migraatiota, mutaatioita tai luonnonvalintaa. (Frankham ym. 2002; Wigginton ym. 2005; Cutter 2019). Alleeli- ja genotyyppifrekvenssit kahden alleelin lokuksissa jakautuvat heterotsygooteilla 0–0.5 välillä ja homotsygooteilla 0–1 Hardyn-Weinbergin tasapainon alla. Useimmiten genotyyppien frekvenssit vastaavat Hardyn-Weinbergin lain odotuksia (Frankham ym. 2002) myös transkriptomi tutkimuksissa (kts. Mastrochirico-Filho ym. 2016; Sun ym. 2018). Poikkeavuudet Hardy-Weinberg tasapainon odotuksista ovat yleensä seurausta sekvenssien duplikaatioista,

genotypointi virheistä, deleetioista, valinnasta tai sukusiittoisuudesta (Wigginton ym. 2005; Chen ym. 2017; Graffelman ym. 2017).

1.2.4. Heterotsygotia

Populaatiogeneettisissä tutkimuksessa heterotsygotiaa käytetään määrittämään populaatioiden yksilöiden geneettistä variaatiota ja fitnessiä. Sukusiittoisuus nostaa populaation homotsygotiaa, mikä laskee sen fitnessiä ja johtaa ajan myötä heterotsygotian menetykseen. Sukusiittoisissa populaatioissa heterotsygotian tarkastelu mahdollistaa toisiin populaatioihin pariutuvien yksilöiden tunnistamisen sukusiittoisista yksilöistä (Allendorf & Leary, 1986; Hansson & Westerberg, 2002).

Havaittu heterotsygotia on heterotsygoottisten yksilöiden määrän keskiarvo populaatiossa sen eri lokuksissa. Odotettu heterotsygotia eli geenidiversiteetti on kahden satunnaisen alleelin todennäköisyys erottautua toisistaan DNA-sekvenssinsä perusteella. Odotetulla heterotsygotialla voidaan mitata, kuinka usein eri yksilöillä esiintyy eri alleeleja, kun yksilöt pariutuvat satunnaisesti. Kahdelle alleelille odotettu heterotsygotia pohjautuu Hardy-Weinberg tasapainon oletukseen heterotsygoottisissa yksilöissä (Cutter, 2019).

1.2.5. Populaatiorakenne

Lajin populaatiorakenne koostuu sen populaatiokoon kasvusta tai pienenemisestä sekä eristyksen ja migraation yhteisvaikutuksesta. Niihin vaikuttavat myös ihmisen aiheuttamat muutokset lajin elinympäristöön. Populaatiorakenteen pohjalta voidaan kuvata populaation yksilöiden geneettistä variaatiota ei-diskreetissä populaatiossa. Sen määrittäminen on tärkeä osa populaatiogenetiikkaa ja siihen on käytetty useimmiten F-statistiikan arvoa F_{ST} kuvaamaan alapopulaatioiden välisiä geneettisiä eroja sekä nykyään enimmissä määrin pääkomponenttianalyysiä (Wright, 1965; Platt ym. 2010; Meirmans & Hedrick, 2011; Meisner & Albrechtsen 2018).

Populaatiorakenteeseen vaikuttavat tekijät, kuten geenivirta, mutaatiot, luonnonvalinta ja ajautuminen ovat tärkeitä yleisellä tasolla ja useissa lajeissa. Populaatioiden geneettiset prosessit ovat haasteellisia erottaa toisistaan yksittäiseen lajiin pohjautuvissa analyyseissä, mutta vertailevilla hypoteeseilla tehdyillä testeillä on mahdollista havainnoida yleisesti näitä prosesseja (Bohonak 1999). Hitaasti leviävissä ja keskenään risteytyvissä lajeissa, jotka ovat levittäytyneet laajalle, esiintyy tyypillisesti populaatioiden eristeisyyttä toisistaan (Platt ym. 2010).

1.2.6. F-statistiikat

Yksi populaatiogenetiikan perimmäisistä osa-alueista on populaatioiden välinen erilaistuminen. Se voidaan laskea heterotsygotialla, jota voidaan käyttää F-statistiikkojen avulla kuvaamaan populaation ja sen alapopulaation geneettistä erilaistumista. Wrightin F-indeksit F_{IS}, F_{ST} ja F_{IT}, missä I = yksilöitä, S = alapopulaatioita ja T = koko populaatio, auttavat mallintamaan geneettistä erilaistumista alapopulaation yksilöiden välillä (F_{IS}), geneettistä erilaistumista alapopulaatioiden välillä (F_{ST}) ja yksilöiden eroja koko populaatiossa (F_{IT}) (Jakobsson ym. 2013; Cutter, 2019).

F_{IS} on sukusiittoisuuskertoimen F:n vastine, joka perustuu useampaan alapopulaatioon. F_{IS} auttaa ymmärtämään miten alapopulaatioissa yksilöt pariutuvat ja siten auttaa hahmottamaan, kuinka paljon eroavaisuutta on Hardy-Weinbergin tasapainon oletukseen. F_{IS} -kaava esitetään alla:

$$F_{IS} = \frac{H_S - H_I}{H_S},$$

missä H_I on havaittu heterotsygotia alapopulaatioissa ja H_S on odotettu heterotsygotia alapopulaatioissa. F_{IS} tulokset voivat olla 1 ja -1 välillä, positiivinen arvo viittaa sukusiittoisuuteen, kun taas 0 viittaa, että sitä ei ole ja negatiivinen arvo viittaa siihen, että yksilöt eroavat enemmän toisistaan kuin odotetaan satunnaisesti pariutuvasta populaatiosta (Cutter, 2019).

F_{ST} kuvaa, kuinka populaation rakenne itsessään vaikuttaa heterotsygotian poikkeavuuksiin odotusarvosta. Se on yleensä F-statistiikoista kiinnostavin biologeille, koska F_{ST} osoittaa, kuinka suuresti populaatiorakenne, migraatio ja elinympäristöjen laajeneminen vaikuttavat geneettiseen variaatioon, ja kuinka geneettisesti erilaistuneita alapopulaatiot ovat verrattuna toisiinsa. Se on myös helpoin statistiikka määrittää, koska toisin kuin F_{IS} tai F_{IT} niin F_{ST}:n laskemiseen tarvitaan vain alleelifrekvenssit genotyyppifrekvenssien sijaan. F_{ST} -kaava esitetään alla:

$$F_{ST} = \frac{H_T - H_S}{H_T},$$

missä H_T on koko metapopulaation odotettu heterotsygotia, joka saadaan yhdistämällä kaikki yksilöt yhteen ottamatta huomioon niiden jakautumista alapopulaatioihin ja H_S on odotettu heterotsygotia alapopulaatioissa. F_{ST} voi saada arvoja 0 ja 1 välillä, missä 0 viittaa siihen, että alapopulaatiot risteytyvät täysin ja 1 viittaa siihen, että alapopulaatioiden välillä on paljon geneettistä variaatiota (Jakobsson ym. 2013; Cutter, 2019).

F_{IT} määrittää, kuinka paljon lajin populaatio eroaa Hardy-Weinberg odotuksista, jos kaikki alapopulaatioiden yksilöt otetaan huomioon. Se kuvaa, mitä tapahtuu, jos populaatioiden jakautumista alapopulaatioihin ei oteta huomioon. F_{IT} -kaava esitetään alla:

$$F_{IT} = \frac{H_T - H_I}{H_T},$$

missä H_T on koko metapopulaation eli kaikkien saman lajin vaihtelevien lähipopulaatioiden odotettu heterotsygotia, joka saadaan yhdistämällä kaikki yksilöt yhteen ottamatta huomioon niiden yksilöryhmiä ja H_I on havaittu heterotsygotia alapopulaatioissa. F_{IT} voi saada arvoja -1 ja +1 välillä, missä positiivinen arvo viittaa sukusiittoisuuteen, kun taas 0 viittaa, että sitä ei ole ja negatiivinen arvo viittaa siihen, että yksilöt eroavat enemmän toisistaan kuin odotetaan satunnaisesti pariutuvasta populaatiosta tapahtuvan (Hanski 1998; Cutter, 2019).

1.2.7. Alleelifrekvenssispektri

Cutter (2019) kuvaa kirjassaan, että nukleotidipolymorfia voidaan esittää käyttämällä niiden frekvenssiä populaatiossa. Tämä tehdään mittaamalla polymorfisten paikkojen harvinaisia alleeleja aloittamalla singletoneista eli polymorfisista paikkoista, jotka esiintyvät vain kerran, ja etenemällä frekvensseissä, kunnes kaikki polymorfiset paikat on käyty läpi. Tätä kutsutaan alleelifrekvenssispektriksi, jota voidaan käyttää määrittämään populaation rakennetta, kuten esimerkiksi populaation pullonkauloja, mistä kertyy tavallista enemmän yleisiä alleeleja, tai populaation kasvua uusien mutaatioiden myötä, kun harvinaisia alleeleja on paljon (Marth ym. 2004; Linck & Battey, 2019).

Harvinaisten alleelien frekvenssi populaatiossa on rajoittunut ja harvinaiset alleelit voivat saada arvoja 0 ja 50 prosentin välillä. Alleelifrekvenssispektriä käytetään kuvaamaan, kuinka paljon alleeleja esiintyy populaatiossa. Tuloksissa odotetaan olevan enemmän polymorfisia paikkoja, joilla on pieni frekvenssi (singleton) ja korkeamman frekvenssin arvoja odotetaan vähemmän. Tyypillisesti, kun alleelifrekvenssi koostetaan ilman referenssiä, siitä lasketaan kuinka monella yksilöllä *i* tai yksilöllä *n – i* esiintyy polymorfinen paikka, mikä puolittaa spektrin ja siten taittaa sen itseensä (Bustamante ym. 2002; Cutter, 2019).

1.3. Tutkimuskysymykset

Tämän pro gradu -tutkielman tarkoituksena on selvittää, kuinka suodattamalla sekvenssidataa voidaan vaikuttaa tulosten laatuun ja datan soveltuvuuteen populaatiogeneettisiin analyyseihin. Sen lisäksi tarkoituksena on selvittää kuinka transkriptomidata soveltuu populaatiogeneettiseen analyysiin.

Lampyridae -heimon tulikärpäsiä ja kiiltomatoja on tutkittu pääasiassa parinvalinnan ja bioluminesenssin näkökannoista. Aikaisempaa populaatiogeneettistä tutkimusta ei ole tehty *L. noctiluca* -lajille. Eroten muista tulikärpäsistä *L. noctiluca* -naaraat kykenevät loistamaan, mikä on johtanut siihen, että suuri osa *L. noctiluca* -lajiin liittyvistä tutkimuksista on pääasiallisesti pohjautunut parinvalintaan. Viljakainen ym. 2020 julkaisemassa tutkimuksessa tutkittiin *L. noctiluca* -lajin RNA-viruksia, minkä aineistosta tässä tutkimuksessa käytetään vain kahden suomalaisen populaation transkriptomia. Tässä tutkielmassa katsotaan näiden populaatioiden välistä ja sisäistä geneettistä muuntelua, kun tiedetään, että naaraat eivät muuta kauas synnyin paikaltaan ja koiraat eivät ole hyviä lentäjiä. Pääoletuksena on, että populaatioiden välillä on erilaistumista.

Tutkimuskysymykseni ovat seuraavat;

- Voidaanko suodatuksella saada data soveltumaan populaatiogeneettisiä analyyseja varten?
- Millaista geneettistä erilaistumista kahden suomalaisen kiiltomato populaatioiden välillä on?

2. Aineisto ja menetelmät

2.1. Näytteenotto

Käytössä oleva aineisto on Viljakainen ym. 2020 tutkimuksessa käyttämän vuoden 2017 *Lampyris noctiluca* kohortin naaraita ja koiraita. Aineiston keräys, eristys ja sekvensointi on kuvattu kyseisessä tutkimuksessa ja tässä tutkimuksessa käytetään vain Hangon ja Konneveden populaatioista kerättyjä yksilöitä. Hangon populaatiosta (N62° 37′, E26° 20′) kerättiin kahdeksan naarasta ja kahdeksan koirasta heinä- ja elokuussa 2017 ja Konneveden populaatiosta (N59° 53′, E23° 06′) kerättiin yhdeksän naarasta ja neljä koirasta kesä- ja heinäkuussa 2017.

Naaraiden takaruumiin valoelimet ja koiraiden päät leikattiin erikseen ja niistä eristettiin RNA käyttäen RNeasy Micro Kittiä (Qiagen). RNA näytteet lähetettiin BGI Tech Solutionsille kirjaston esikäsittelyyn käyttäen Illumina TruSeq Stranded mRNA Library Prep Kittiä ja RS-122–2101 kittiä. Tehdyt kirjastot sekvensoitiin Illumina HiSeq4000 PE100:lla, josta saatiin parittaiset sekvenssilukemat. Sekvensoiduista raakalukemista poistettiin adapterisekvenssit, laadultaan huonot (Q < 20) nukleotidit ja alle 36 nukleotidia lyhyemmät lukemat. Trimmatut lukemat yhdistettiin ja koostettiin Trinity v2.3.2 (Grabherr ym. 2011) -ohjelmalla *in-silico* normalisaatiolla (Viljakainen ym. 2020).

Laadun tarkastus tehty Beijing Genome Institution (BGI) puolesta Agilent 2100 bioanalysaattorilla. Totaali-RNA:ta lähetettiin BGI:lle 15 µl, josta on saatu raportit, joista voidaan todeta, että laadullisesti näytteet ovat hyväksyttävissä rajoissa (Q20 >98 %) ja yhtä näytettä lukuun ottamatta RIN-arvo on > 6, poikkeavassa näytteessä RIN-arvo on 2.6. GC (%) -pitoisuus kaikissa näytteissä on 38 %:sta 40 %:n. Kaikki näytteet lukuun ottamatta yhtä yksilöä todettiin BGI:n toimesta kelvollisiksi, mutta Hangon populaation näyte MS02 todettiin riskialttiiksi.

2.2. BUSCO

Sekvensoinnista tuotetun genomisen datan määrän kasvaessa sen laadun arvioinnista on tullut myös entistä tärkeämpää. Kuitenkaan laadunarvioinnin menetelmät, jotka pohjautuvat k-meereihin ja contigien pituuksiin, eivät kykene arvioimaan sekvensoidun genomin tai transkriptomin geenisisällön kokonaisuutta. BUSCOt ovat yleisiä vertailuun suunniteltuja yhden kopion ortologisia geenejä, joita käytetään arvioimaan sekvensoinnin kokonaisuutta, koska evoluution näkökulmasta on todennäköistä, että sekvensoiduille geeneille on olemassa ainakin yksittäinen ortologinen kopio. BUSCO-analyysi mahdollistaa genomin tai transkriptomin kokonaisuuden määrittämisen, mistä saadaan tuloksena selvitettyä geenisisällöstä tutkitun genomiin tai transkriptomin sisältämät kokonaiset geenit, duplikaatit ja mahdollisen puuttuvan datan (Simão ym. 2015).

2.3. Kallisto

Transkriptitason RNASeq-analyyseissä ensimmäiset vaiheet ovat linjaaminen joko referenssigenomiin tai transkriptomiin ja saadun datan transkriptien määrän mittaaminen. Transkriptien määrän määrittäminen on tyypillisesti aikaa vievä prosessi, minkä nopeuttamiseksi Kallisto-ohjelma perustuu pseudolinjauksien käyttämiseen tyypillisen linjauksen sijaan. Kallisto tekee tämän tunnistamalla, mistä transkriptista lukemat ovat peräisin eikä mihin sekvensseihin

ne linjautuvat. Kallisto hyödyntää k-meerejä nopeasti hajauttamalla ne yhteen transkriptomipohjaisen *de Brujin* kaavion (T-DBG) kanssa, jotta lukemista saadaan tarkka pseudolinjaus transkriptomiin (Bray ym. 2016).

Kallisto käyttää transkriptomi *de Brujin* kaaviota (T-DBG) ja sen reititystä, joka koostuu reiteistä, joiden yhteenliittymä peittää kaavion reunat, missä reitit vastaavat transkriptejä. T-DBG:n peittävät reitit indusoivat k-yhteensopivuusluokkia. Yhteensopivuusluokka voidaan määrittää virheettömään lukemaan esittämällä se reittinä kuvaajassa ja määrittämällä se k-yhteensopivuusluokan k-meerin ja sen k-yhteensopivuusluokkien risteyskohtana. Lukeman ekvivalenttiluokka koostuu siihen yhdistetyistä transkripteistä, joista lukema on mahdollisesti peräisin, ja siten antaa tarvittavan statistiikan transkriptien määrän kvantifiointiin (Bray ym. 2016).

2.4. N50 ja ExN50

N50-contig saadaan laskemalla contigien pituudet yhteen pisimmästä lyhimpään contigiin, kunnes ensimmäinen pitkä contig ylittää 50 % määrän koko genomin koosta. Tyypillisesti N50-contigia käytetään kuvaamaan genomikokoonpanon laatua. Vastaavasti N10 - N100 vastaavat 10–100 % koko genomin koosta. Seurauksena contigien laskutavasta pienemmän Nx-statistiikan kasvaessa contigin pituus lyhenee. Vaihtoehtona N50-statistiikalle transkriptomitutkimuksessa on ExN50-statistiikka, joka lasketaan samalla tavoin kuin N50, mutta rajoittuu korkeasti ekspressoituneisiin geeneihin, jotka vastaavat x % normalisoidusta ekspressiosta.

2.5. CD-HIT

Sekvenssianalyysi on uuden sukupolven sekvensoinnin myötä entistä tärkeämpi osa bioinformatiikkaa ja aineistojen määrän kasvaessa redundanssin poistaminen datasta on tärkeää. CD-HIT on päällekkäisyyksien poistamiseen käytetty algoritmi, joka vertaa syötettyjä sekvenssejä genomiin aloittaen pisimmästä sekvenssistä ja siirtyen lyhyimpään. Ensimmäinen tunnistettu sekvenssi määritetään edustavaksi sekvenssiksi, mihin seuraavia sekvenssejä verrataan. Sekvenssi tunnistetaan päällekkäiseksi tai edustavaksi sekvenssiksi riippuen siitä, kuinka samankaltainen se on verrattuna muihin ennen sitä tunnistettuihin edustaviin sekvensseihin (Fu ym. 2012).

Samankaltaisuus perustuu sananlaskenta- ja indeksointitaulukkoon, joilla suodatetaan ylimääräiset verrokit. Jokaista käsiteltävää sekvenssin nukleotidia verrataan edustaviin

sekvensseihin, ja yhtenevien nukleotidien määrä käsiteltävästä sekvenssistä ja noudetuista edustavista sekvensseistä päivitetään, kunnes sekvenssit on käyty läpi. Tämän jälkeen edustavat sekvenssit, jotka ylittävät määritetyn samankaltaisuuden rajan, linjataan käsiteltävään sekvenssiin samankaltaisuuden määrittämiseksi (Fu ym. 2012).

2.6. Sekvenssilukemien linjaus

Sekä Hangon että Konneveden populaatioista koostetut transkriptomit yhdistettiin molemmille populaatioille yhteiseksi yhdistetyksi transkriptomiksi cat-komennolla, jotta populaatioiden lukemat saadaan linjattua yhdistettyyn transkriptomiin varianttien tunnistusta varten STAR-aligner -ohjelmalla (Dobin ym. 2012). Tälle yhdistetylle transkriptomille tehtiin CD-HIT-käsittely (Fu ym. 2012) poistamaan redundanssia sekvensseistä käyttäen CD-HIT-Est -toimintoa parametreilla (-c 0.98 -M 1000) eli sekvenssien samankaltaisuuden tulee olla 98 % ja käytetty muisti on 1000 megabittiä.

Näytteiden reverse ja forward sekvenssilukemat linjattiin STAR-aligner ohjelmalla CD-HIT-käsiteltyyn, yhdistettyyn transkriptomiin käyttäen parametria --outFilterMismatchNoverLmax 0.05, mikä rajoittaa yhteensopimattomat linjaukset 0.5 prosenttiin koko sekvenssin pituudesta. STAR-alignerin ajosta saatiin jokaiselle näytteelle linjattu bam-tiedosto eli pakattu sekvenssilinjauskartta, joka sisältää referenssiin linjatun sekvenssidatan, jota käytetään varianttien tunnistukseen.

2.7. Varianttien tunnistus

RNAseq-datassa olevien varianttien tunnistuksessa noudatettiin GATK 4.0 (Poplin ym. 2017) -ohjelmiston ohjeistusta (https://gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNAseq-short-variant-discovery-SNPs-Indels-). Koska referenssigenomi puuttui, Base Quality Recalibration vaihe jätettiin tekemättä. Ennen varianttien tunnistusta STAR-aligner ohjelmalla linjatuista bam-tiedostoista koostettiin Picardin RevertSam toiminnolla linjaamattomat bam-tiedostot, joista lukuryhmien eli sekvensointivälineen ajon tiedot saatiin yhdistettyä dataan Picardin FastqToSam- ja MergeBamAlign-toiminnoilla. Varianttien tunnistus tehtiin jokaiselle yksilölle erikseen. Tämän jälkeen jokaisen yksilön VCF-tiedosto, joka sisältää yksilön geenisekvenssivariaatiodatan, yhdistettiin Konneveden ja Hangon populaatioiden tiedostoiksi populaatioiden sisäisen muuntelun tutkimista varten sekä molemmista populaatioista yhdistetyksi yhteistiedostoksi, joka jaettiin naaras- ja koirasnäytteisiin populaatioiden välisen muuntelun tutkimista varten, bcftools-ohjelmalla.

Varianttien suodatus tehtiin GATK 4.0 -ohjelmistolla yhdistetyille VCF-tiedostoille ja niistä poistettiin sellaiset sekvenssit, joiden sekvensoinnissa reverse- ja forward-juosteissa toista juostetta on suosittu (SOR > 3.0 ja FS > 60.0). Samoin poistettiin huonosti kartoitetut lukemat, joiden laadun neliöllinen keskiarvo on enemmän kuin 40, kartoituksen laadun u-testin arvo on pienempi kuin -12.5 ja alleelipaikkojen u-testin arvo on pienempi kuin -8.0 (MQ < 40.0, MQRankSum < -12.5 ja ReadPosRankSum < -8.0). Sekvenssien suodatuksen jälkeen transkripteja verrattiin virustietokantaan BLAST-ohjelmalla, minkä jälkeen virusperäiset sekvenssit poistettiin GATK 4.0 -ohjelmalla SelectVariants-komennolla. Tämän jälkeen tiedostoista suodatettiin VCFtools -ohjelmalla (Danecek ym. 2011) lokukset, joissa on huonolaatuiset sekvenssit eli joiden vähimmäislukusyvyys jäi alle 10 ja vähimmäislaatu jäi alle 20 sekä yli puolet on puuttuva dataa (--max-missing 0.5) ja lokukset, joissa on enemmän ja vähemmän kuin kaksi alleelia (--min-alleles 2 --max-alleles 2).

2.8. Populaatiogeneettiset analyysit

VCF-tiedostot luettiin R-ohjelmistoon käyttäen R-pakettia VCFR (Knaus & Grünwald, 2017), minkä jälkeen poistettiin yksilöt, joiden SNP:ssä on enemmän kuin 50 % puuttuvaa dataa per yksilö. Koirasnäytteistä poistettiin Hangon populaatiosta yksilö MS02, joka oli aikaisemmin tunnistettu BGI:n avulla riskialttiiksi ja siitä havaittiin yli 50 % puuttuvaa dataa.

Populaatiogeneettiset analyysit tehtiin käyttäen R-paketteja adegenet (Jombart, 2008) ja pegas (Paradis, 2010). Analyysien tulosten kuvaajat tehtiin R-paketilla ggplot2 (Wickham, 2016). Samalla jokaisesta VCF-tiedostosta tehtiin alleelifrekvenssispektrit pegas R-paketilla tunnistamaan populaatioiden alleelivariaatiota.

2.9. McKinneyn menetelmä

Tutkimuksissa, joissa sekvensointi on tehty RADSeq- tai RNASeq-menetelmällä (Rellstab ym. 2019), suurena ongelmana on paralogien tunnistaminen sekvenssidatasta niiden sekvenssien samankaltaisuuden ja pituuden takia, minkä seurauksena paralogit yleensä poistetaan ennen analyysejä. Alleelifrekvensseihin pohjautuvissa tutkimuksissa paralogit ovat ongelmallisia, koska niiden samankaltaisuus usein johtaa niiden tunnistamiseen samaksi sekvenssiksi, mikä vääristää alleelifrekvenssejä ja vaikeuttaa genotypointia (McKinney ym. 2017; McKinney ym. 2018; Rellstab ym. 2019).

McKinneyn menetelmä perustuu paralogien tunnistukseen luonnollisista populaatioista kahden tekijän pohjalta: populaatioiden heterotsygoottisten yksilöiden määrän ja niiden sisäisen alleelisuhdeluvun. Yhdistettäessä heterotsygoottien määrän (H) populaatiossa jokaiselle lokukselle ominaisiin alleelilukemiin 1:1 suhteessa (D) mahdollistaa paralogisten lokusten tunnistuksen kaikista alleelifrekvensseistä (McKinney ym. 2017).

Heterotsygoottisten yksilöiden (H) odotettu lukumäärä tulee olla suurempi lokuksille, joista on useita kopioita kuin yksittäisille lokuksille. Niinpä heterotsygoottien lukumäärä mahdollistaa duplikaattien ja yksittäisten lokusten tunnistamisen toisistaan, kun niiden frekvenssit ovat tasapainossa. Heterotsygoottisten yksilöiden alleelilukemat ovat myös erilaisia yksittäisten lokusten ja duplikaattien välillä. Yksittäisten lokusten genotyypit ovat heterotsygoottisissa yksilöissä symmetrisessä 1:1 suhteessa, mutta duplikaateilla heterotsygoottiset genotyypit ovat epäsymmetrisessä 1:3 tai 3:1 suhteessa. Yksilöt, joiden genotyypit ovat epäsymmetriassa, odotetaan lähestyvän epäsymmetristä alleelisuhdetta, mikä mahdollistaa duplikaattilokusten tunnistamisen. Paralogisia lokuksia voidaan tunnistaa kaikista alleelifrekvensseistä populaation heterotsygoottien lukumäärän ja lukemasuhde eroavaisuus (D) -statistiikan avulla (McKinney ym. 2017).

Kaikista VCF-tiedostoista tehtiin McKinneyn menetelmällä kuvaajat käyttäen R-pakettia HDplot (McKinney ym. 2017), joiden pohjalta yhdistetystä transkriptomista tehtiin kolme VCF-tiedostoa, jotka suodatettiin seuraavasti: Ensimmäisessä suodatettiin lokukset, joissa McKinneyn lukusuhdepoikkeama (D) on enemmän kuin 20 tai vähemmän kuin -20. Toisessa suodatettiin lokukset, joissa odotettu heterotsygotia (H) on enemmän kuin 0.6. Viimeisessä tiedostossa tehtiin kummatkin suodatukset. Hangon ja Konneveden näytteille sekä yhdistetyn populaation koiras- ja naarasnäytteille tehtiin McKinneyn menetelmää käyttäen suodatus arvoilla $-20 < D < 20$ ja $H < 0.6$.

2.10. PCA

Pääkomponenttianalyysi (PCA) on datan dimensioredusointiin perustuva lähestymistapa datan käsittelyyn. Sillä etsitään alkuperäisestä aineistosta lineaarisia yhdistelmiä, joita kutsutaan pääkomponenteiksi (PC), ja joilla on mahdollista havainnoida mittaustulosten vaikutusta. Pääkomponentit ovat kohtisuoraan vastakkaiset keskenään ja niiden ulottuvuudet saattavat poiketa alkuperäisistä arvoista redusoinnin seurauksena. PCA tehdään käyttämällä korrelaatiomatriiseja, joten sen oletetaan teoreettisen validoinnin seurauksena noudattavan normaalijakaumaa (Ma & Dai, 2011).

Bioinformaattisissa data-analyyseissä aineiston korkeat dimensiovaihtelut tuovat omat haasteensa (Ma & Dai, 2011). Populaatiogenetiikassa pääkomponenttianalyysia käytetään populaatioiden ja näytteiden välisen perimän eroavaisuuksien tutkimiseen riippumatta niiden populaatiorakennetta ympäröivistä historiallisista malleista. Pääasiassa kun ensimmäisen pääkomponentin varianssin määrä on suuri, populaatioiden alarakennetta voidaan määrittää ja tilastollisesti merkitsevät pääkomponentit koko aineistosta voidaan tunnistaa (Reich ym. 2008).

McKinney suodatetuille VCF-tiedostoille tehtiin populaatiogeneettiset analyysit (heterotsygotia, F-statistiikat ja alleelifrekvenssispektrit) ja niiden visualisoinnit. Samalla tehtiin pääkomponenttianalyysit koko yhdistetyn populaation datalle ja sen naaras- ja koirasnäytteille erikseen sekä McKinney suodatetuille yhdistetyn populaation näytteille että naaras- ja koirasnäytteille R-paketilla ade4 (Drey & Dufour, 2007)

3. Tulokset

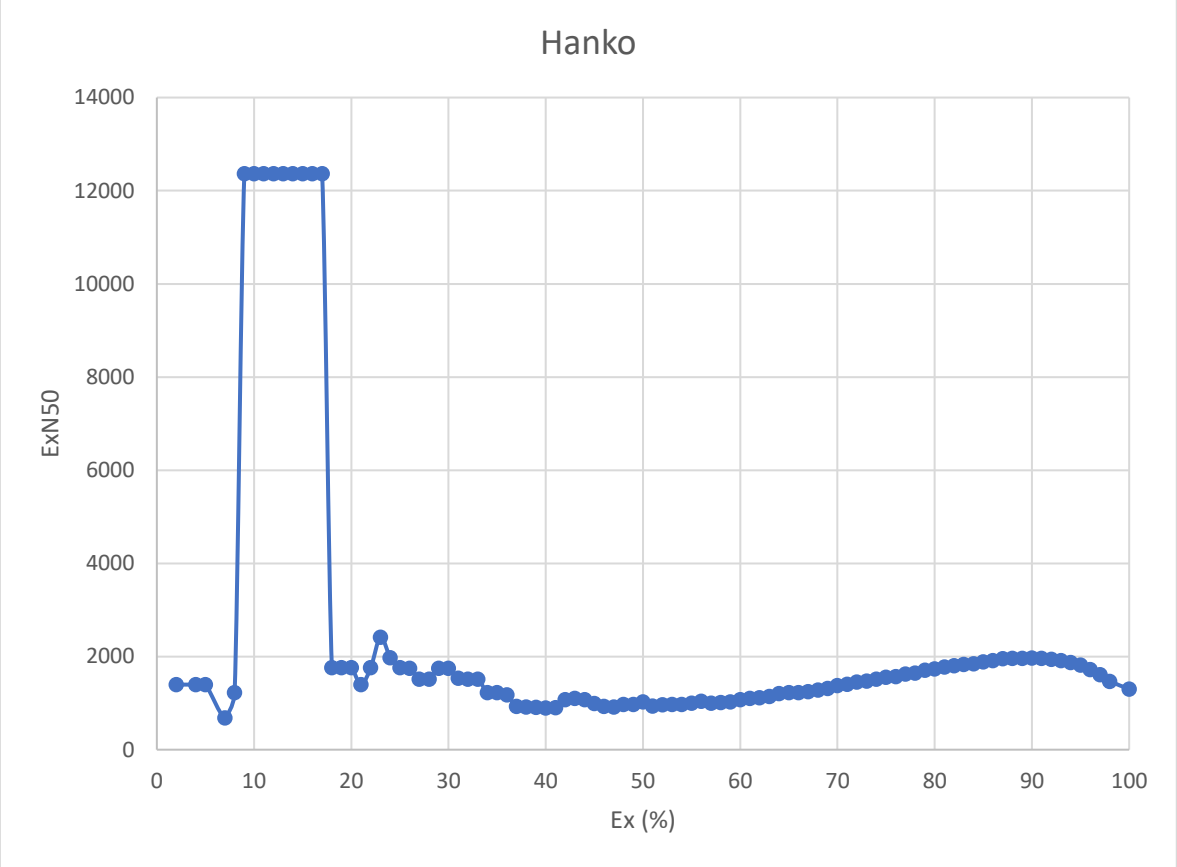
3.1. Transkriptomi-aineisto

Kummankin populaation transkriptomille tehtiin Kallisto-analyysit transkriptien laadun ja mahdollisten ongelmien tunnistusta varten ennen muita analyysejä ja suodattamista. Hangon populaatiossa oli enemmän transkripteja, mutta CG-pitoisuudessa ei ollut suurta eroa (Taulukko 1). Hangon populaatioin ekspressoitu N50-contigin pituuden kuvaaja teki korkean nousun Ex10 – Ex20 ekspressioiden välillä (Kuva 1) ja Konnevedellä Ex5 – Ex15 välillä (Kuva 2). Virusten poistaminen transkriptomeista ei muuttanut kuvaajia, vaikka osassa näytteitä oli paljon viruksia ja havaittujen virusgenomien koko vaihteli 1400–19000 emäsparin välillä (Viljakainen ym. 2020). Siispä on todennäköistä, että kuvaajien piikki johtuu duplikaateista, koska CD-HIT käsittelyn jälkeen kuvaajat muuttuvat (kts. CD-HIT).

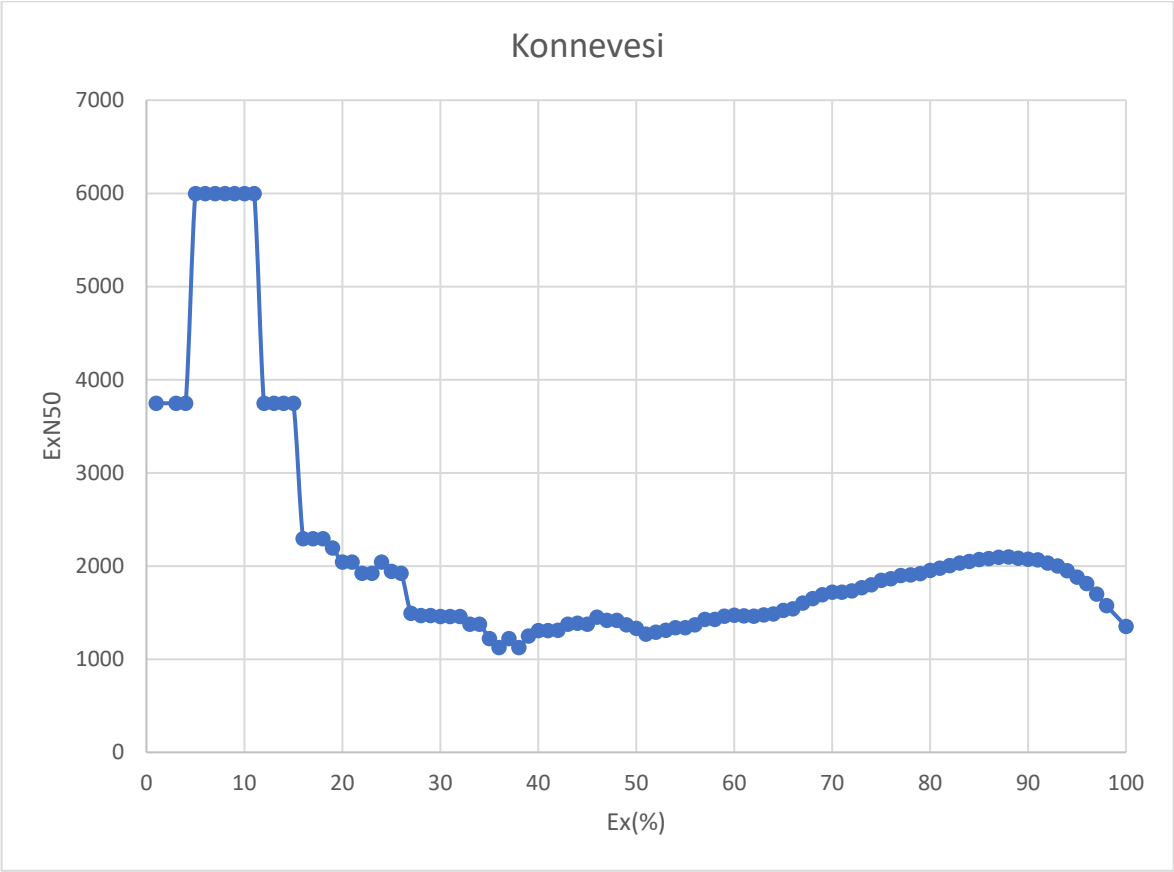
Hangossa parhaat ExN50-contig arvot olivat E23N50-contigin 2416 emäsparia ja E91N50-contigin 1967 emäsparia, mutta E23N50-contigin arvo on mahdollinen poikkeama. Vastaavasti Konnevedellä E89N50 sai korkeimman contigin pituuden 2086, joka on suurempi kuin Hangossa vastaten N50-contigin arvoa.

Taulukko 1 - Hangon ja Konneveden Kallisto-analyysin Trinity-geenien ja transkriptien lukumäärä sekä GC-pitoisuus.

Transkriptien ja ’geenien’ lukumäärä		
Populaatio	Hanko	Konnevesi
Kaikki Trinity ’geenit’	203859	195561
Kaikki Trinity transkriptit	455886	429584
GC %	36.11	36.10



Kuva 1 – Hangon ekspressoidun N50-contigin pituuden kuvaaja. X-akselilla on prosenttiosuudet, jotka edustavat tiettyä prosentuaalista osuutta koko aineiston normalisoidusta ekspressiosta rajoittuen jokaisessa prosenttiluokassa korkeimmin ekspressoituihin geeneihin. Y-akselilla on contigin ekspressoitu N50-pituus, jota merkitään ExN50, jossa x indikoi tiettyä prosenttiosuutta.



Kuva 2 - Konneveden ekspressoitu N50-contigin pituuden kuvaaja. X-akselilla on prosenttiosuudet, jotka edustavat tiettyä prosentuaalista osuutta koko aineiston normalisoidusta ekspressiosta rajoittuen jokaisessa prosenttiluokassa korkeimmin ekspressoituihin geeneihin. Y-akselilla on contigin ekspressoitu N50-pituus, jota merkitään ExN50, jossa x indikoi tiettyä prosenttiosuutta.

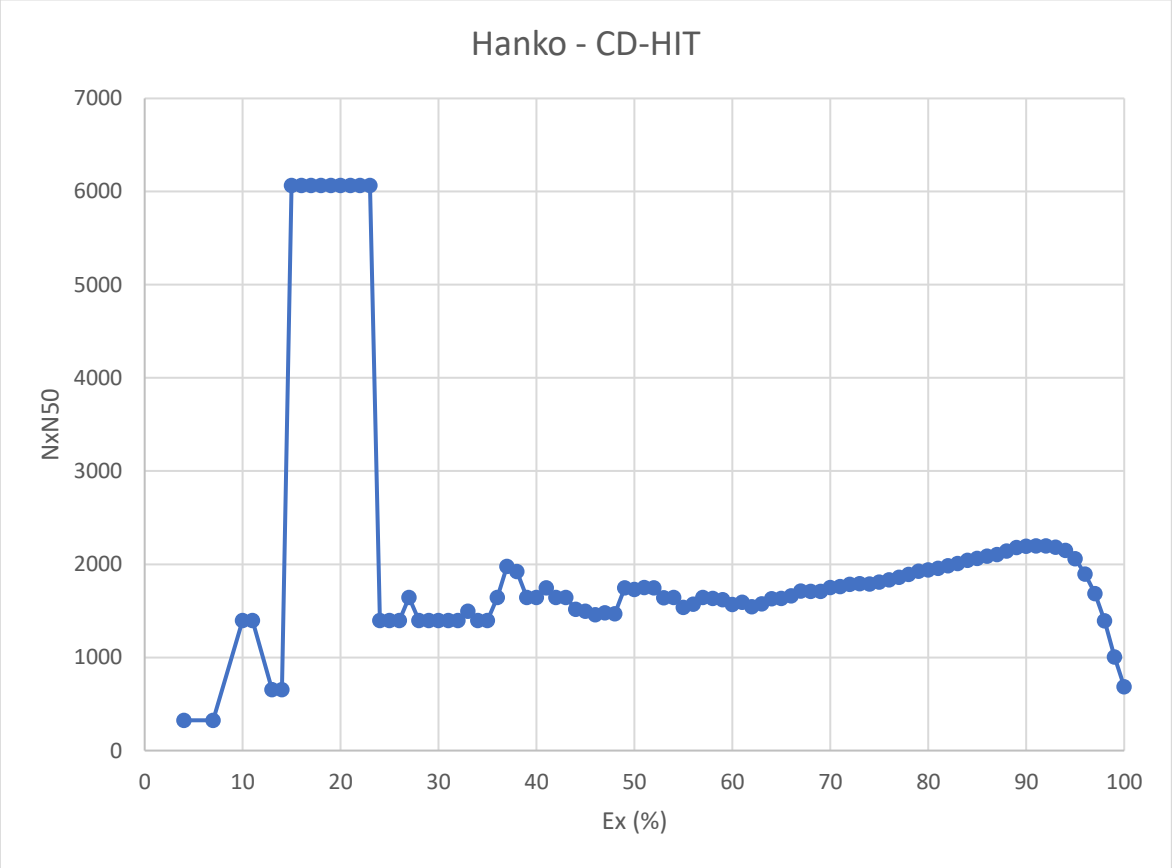
3.1.1. CD-HIT

CD-HIT käsittely tehtiin yhdistetylle transkriptomille. Yhdistetyn transkriptomin tulokset esitetään taulukossa 2, missä keskiarvo contig ja contigien pituus oli suurempi kuin Hangon ja Konneveden populaatiossa lukuun ottamatta N10-contigia. Transkriptien määrä muuttui CD-HIT käsittelyn myötä 455886 transkriptistä 367262 transkriptiin Hangossa ja 429584 Konneveden transkriptistä 340556 transkriptiin. Ennen CD-HIT käsittelyä ekspression kuvaajissa 1–2 havaittu kasvupiikki puolittui Hangossa, mutta pysyi ennallaan Konnevedellä CD-HIT käsittelyn jälkeen (Kuva 3–4). Kuitenkin Hangossa prosenttiosuuden väli siirtyi Ex10 – Ex20 väliltä välille Ex15 – Ex23 ja vastaavasti Konnevedellä Ex5 – Ex15 väliltä välille Ex5 – Ex9. Muutokset kuvaajissa osoittaa, että transkriptomeista esiintyy päällekkäisyyttä. Kasvupiikin alueen contigeista katsottiin annotaatiot, mistä huomattiin, että tämän alueen contigit pääasiassa vastaavat lipidien, fosfaattien ja proteiinien sidonnasta ja kuljetuksesta. Hangon ja Konneveden suurin contig kasvupiikin alueella vastaa vitallogeniiniä, joka vastaa naarashyönteisillä ruskuaisen esiastetta.

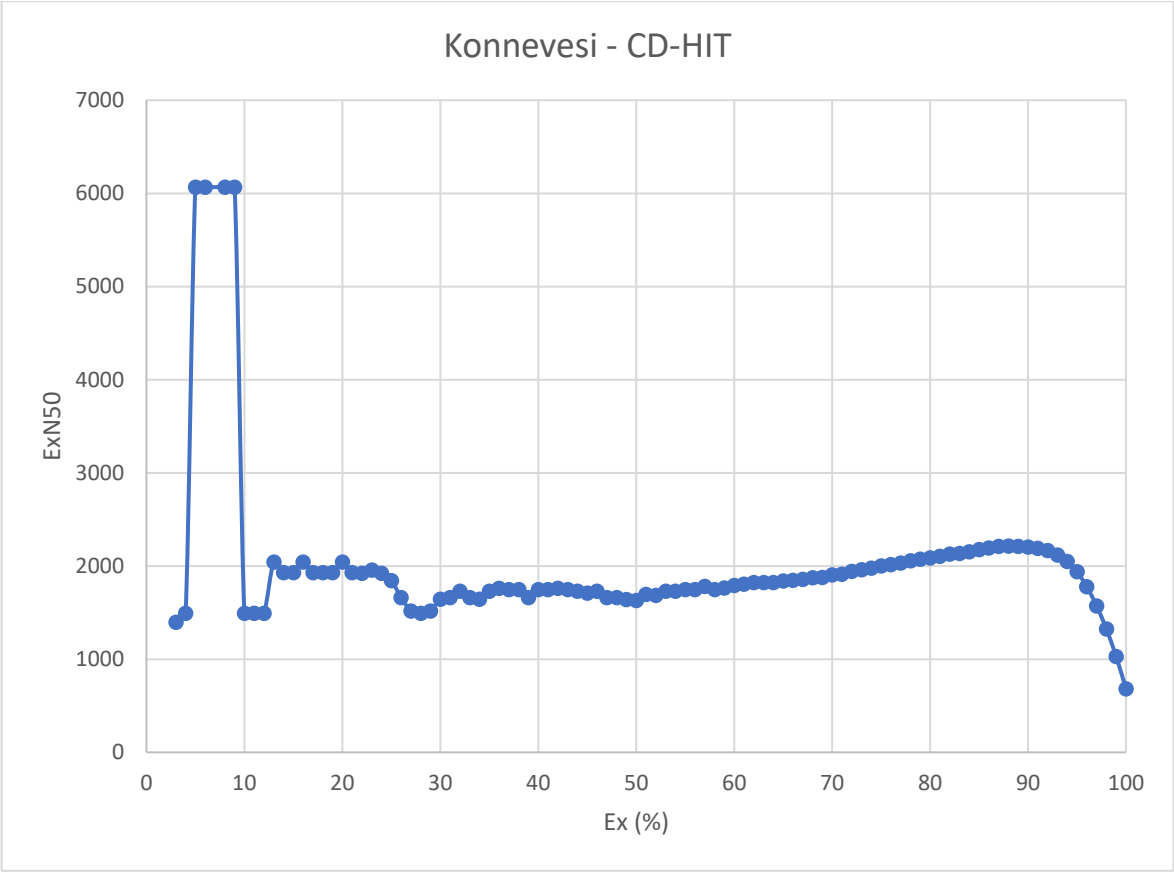
Kummassakin populaatiossa ExN50-contigien korkein emäsparien määrä kasvoi CD-HIT käsittelyn jälkeen saaden pisimmän arvon Hangossa E91N50-contigissa 2197 emäsparilla ja Konnevedellä E88N50-contigissa 2215 emäsparilla.

Taulukko 2 - Hangon ja Konneveden Kallisto-analyysin Trinity-geenien ja transkriptien lukumäärä sekä GC-pitoisuus CD-HIT käsittelyn jälkeen.

Transkriptien ja 'geenien' lukumäärä CD-HIT-käsittelyn jälkeen		
Populaatio	Hanko	Konnevesi
Kaikki Trinity 'geenit'	80519	169169
Kaikki Trinity transkriptit	367262	340556
GC %	36.17	36.16



Kuva 3 - Hangon ekspressoitu N50-contigin pituuden kuvaaja CD-HIT käsittelyn jälkeen. X-akselilla on prosenttiosuudet, jotka edustavat tiettyä prosentuaalista osuutta koko aineiston normalisoidusta ekspressiosta rajoittuen jokaisessa prosenttiluokassa korkeimmin ekspressoituihin geeneihin. Y-akselilla on contigin ekspressoitu N50-pituus, jota merkitään ExN50, jossa x indikoi tiettyä prosenttiosuutta.



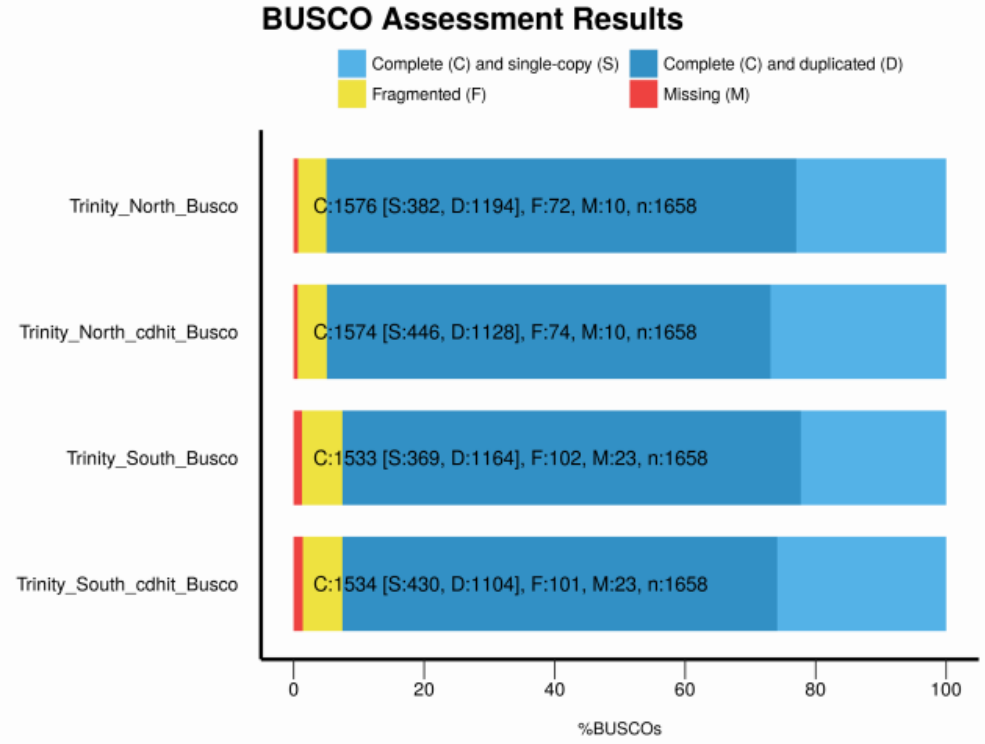
Kuva 4 - Konneveden ekspressoitu N50-contigin pituuden kuvaaja CD-HIT käsittelyn jälkeen. X-akselilla on prosenttiosuudet, jotka edustavat tiettyä prosentuaalista osuutta koko aineiston normalisoidusta ekspressiosta rajoittuen jokaisessa prosenttiluokassa korkeimmin ekspressoituihin geeneihin. Y-akselilla on contigin ekspressoitu N50-pituus, jota merkitään ExN50, jossa x indikoi tiettyä prosenttiosuutta.

3.1.2. BUSCO

Ennen tätä tutkimusta tehdyssä BUSCO-analyysistä saatiin selville, että Hangon ja Konneveden populaatioista transkriptomit ovat hyvälaatuisia, mistä vain 5–10 % BUSCO:ista puuttuu tai ovat sirpaloituneita (Taulukko 3). Konneveden populaatio oli verrattuna Hangon populaatioon kokonaisempi, mutta kuitenkin kummassakin populaatiossa oli suuri määrä duplikaatti BUSCO:ja, mikä viittaa suureen määrään duplikaattigeenejä (Kuva 5).

Taulukko 3 – Yhteenveto BUSCO hausta, jossa verrattiin 1658 yhdenkopion ortologia lahkosta insecta, tuloksista, jotka luokitellaan kokonaisiin, duplikoituneihin, sirpaloituneihin ja puuttuviin ortologeihin ennen ja jälkeen CD-HIT-käsittelyn. Taulukossa esitetään, kuinka moni transkripti vastaa yhdenkopion ortologiin.

BUSCO	Hanko		Konnevesi	
	Alkuperäinen	CD-HIT	Alkuperäinen	CD-HIT
Kokonaiset BUSCOt (C)	1533	1534	1576	1574
Kokonaiset yhdenkopion BUSCOt (S)	369	430	382	446
Kokonaiset duplikaatti BUSCOt (D)	1164	1104	1194	1128
Sirpaloituneet BUSCOt (F)	102	101	72	74
Puuttuvat BUSCOt (M)	23	23	10	10
Kaikki tutkitut BUSCO-ryhmät	1658	1658	1658	1658



Kuva 5 – BUSCO tulosten pohjalta koostettu kerrostettu pylväsdiagrammi kummastakin Konneveden (North) ja Hangon (South) transkriptomeista. Tulokset on esitetty ennen CD-HIT-käsittelyä ja sen jälkeen. CD-HIT-käsitellyissä transkriptomeissa näkyy lievä kasvu kokonaisten yhdenkopion BUSCOjen määrässä.

3.1.3. Linjautuminen

STAR-alignerilla linjattiin sekvenssilukemat populaatiokohtaisista transkriptomista koostettuun yhdistettyyn transkriptomiin. Kummankin populaation yksilöt linjautuivat hyvin lukuun ottamatta yksilöitä MS02, MS04 ja MS08, joilla vain 53, 77 ja 65 prosenttia lukemista

linjautui. Suurin osa (60–67 %) kartoitetuista lukemista linjautui useaan kertaan transkriptomiin, mikä mahdollisesti johtui Gayral ym. 2013 julkaisussa esitetystä piilotetusta paralogiasta. Yhteen genomiseen paikkaan linjautuneita lukemia oli tasaisesti noin 23–27 % välillä (Taulukko 4).

Taulukko 4 – STAR-alignerin linjautumisen statistiikat, joista näkyy kuinka suuri osa sekvensseistä linjautuu yhteen tai useampaan paikkaan transkriptomissa, kuinka paljon lukemia per yksilö, sekvenssien keskiarvoiset pituudet STAR-alignerin oman suodatuksen jälkeen ja kuinka moni lukema linjautui.

	Syötetyt lukemat	Keskiarvoinen sekvenssin pituus	Yhteen paikkaan linjautuneet (%)	Moneen paikkaan linjautuneet (%)	Linjautuneet lukemat (%)	Linjautumattomat lukemat (%)
Konnevesi						
FN01	29224023	196.14	23.36	67.48	94.88	5.11
FN02	34312874	196.4	23.22	66.18	94.59	5.42
FN03	35377372	196.17	23.74	66.29	94.39	5.6
FN04	40093367	196.31	25.11	64.51	94.86	5.13
FN05	36660872	195.8	27.82	60.4	92.86	7.15
FN06	41181729	196.06	22.82	67.32	94.66	5.34
FN07	36022939	196.06	24.14	64.64	93.47	6.52
FN08	37923198	195.86	23.95	63.26	92.13	7.87
FN11	37449265	195.6	24.27	65.51	94.28	5.73
MN01	32006641	195.98	22.57	67.27	94.73	5.27
MN02	34271083	195.97	22.9	67.75	95.31	4.7
MN03	41759156	196.06	22.11	67.61	94.47	5.53
MN05	31507216	195.9	23.3	67.04	95.12	4.87
Hanko						
FS01	42370781	196.02	25.88	65.1	94.89	5.11
FS02	34265408	196.12	25.33	65.87	95.20	4.79
FS03	31331434	195.96	24.99	64.93	94.11	5.89
FS04	39038484	195.39	25.84	62.57	92.80	7.19
FS05	29704600	195.81	23	65.59	92.26	7.73
FS06	27078807	195.94	23.97	66.96	94.76	5.23
FS07	35265610	195.79	24.74	65.14	93.69	6.3
FS08	32089459	196.02	25.21	63.48	92.38	7.61
MS01	34082183	196.01	23.75	66.42	95.10	4.89
MS02	27279300	189.58	17.12	33.67	53.15	46.85
MS03	31333099	195.9	23.59	65.21	94.01	5.99
MS04	33483246	190.8	27.35	46.35	77.04	22.97
MS05	36067384	195.92	22.98	66.5	94.38	5.62
MS06	40751912	196.17	23.62	65.88	94.49	5.5
MS07	29802351	196	24.32	64.51	93.90	6.1
MS08	36085449	192.72	19.39	43.09	65.81	34.18

3.2. Transkriptomi-aineiston analyysit ja suodatukset

Tässä osiossa käyn ensiksi läpi transkriptomi-aineiston populaatiogeneettiset analyysit ennen paralogien poistamista McKinneyn-menetelmällä. Sitten käyn läpi aineiston populaatiogeneettiset analyysit paralogien poistamisen jälkeen ja tarkastelen, kuinka niiden suodattaminen vaikuttaa tuloksiin.

3.2.1. Suodatus ja SNP:t per populaatio

Tutkimuspopulaatiot suodatettiin siten, että kaikista populaatioista poistettiin lokukset, joissa oli enemmän kuin 50 % puuttuvaa dataa per yksilö ja per populaatio. Samoin populaatioista suodatettiin duplikoituneet lokukset ja sellaiset lokukset, joissa oli enemmän ja vähemmän kuin kaksi alleelia eli analyysit siis suoritettiin muunteleville kahden alleelin lokuksille.

Yhdistetyn aineistossa suodatuksen jälkeen oli 28 näytettä (9 naarasta + 4 koirasta Konnevedeltä sekä 8 naarasta + 7 koirasta Hangosta) ja 7378 SNP per näyte. Konneveden populaatiossa suodatuksen jälkeen oli 13 näytettä (9 naarasta + 4 koirasta) ja 8486 SNP per näyte. Hangossa populaatiossa suodatuksen jälkeen oli 15 näytettä (8 naarasta + 7 koirasta) ja 8502 SNP per näyte. Yhdistetyn aineiston naarasnäytteitä suodatuksen jälkeen oli 17 näytettä ja 8004 SNP per näyte. Yhdistetyn aineiston koirasnäytteitä suodatuksen jälkeen oli 11 näytettä ja 10288 SNP per näyte.

3.2.2. Heterotsygotia

Hangon ja Konneveden populaatioissa laskettiin odotettu ja havaittu heterotsygotia. Odotettu heterotsygotia vaihteli 0.36–0.37 välillä ja vastaavasti havaittu heterotsygotia vaihteli 0.54–0.56 välillä. (Taulukko 5). Hangon populaatiossa oli vähän korkeampi odotettu ja havaittu heterotsygotia.

Taulukko 5 - Keskiarvot Hangon ja Konneveden populaatioiden heterotsygotioista

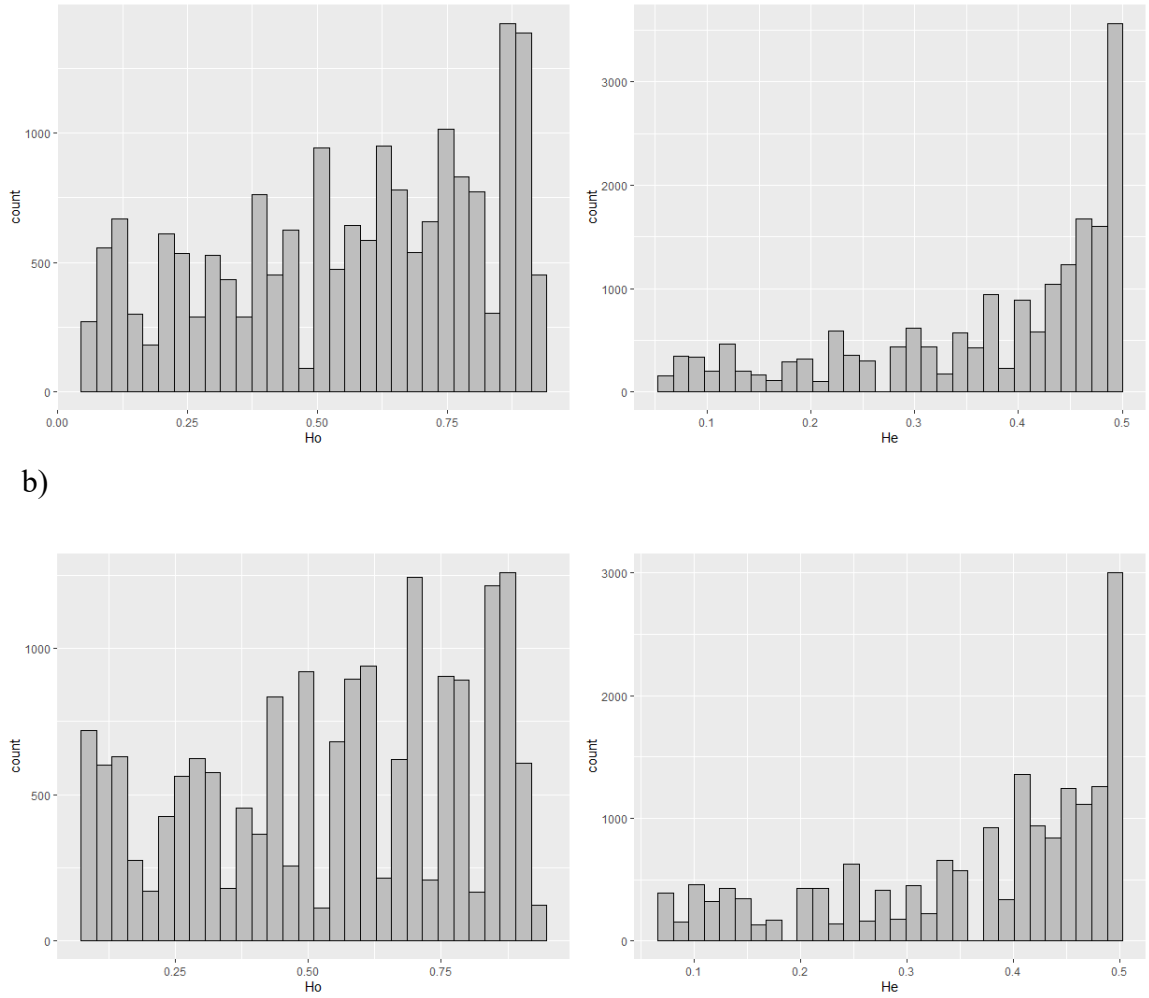
Populaatio	Ho	He
Hanko	0.5647	0.3726
Konnevesi	0.5479	0.3658

Katsottaessa koko aineistoa (Kuva 6) huomattiin, että kohtalaisen suuri osa odotetun heterotsygotian arvoista oli lähellä 0.5, mikä merkitsee sitä, että suuri osa populaatiosta oli

heterotsygootteja. Vastaavasti kummankin populaation havaittu heterotsygotia sai korkeita arvoja varsinkin Konnevedellä (Kuva 6b). Kuitenkin Konneveden tuloksista oli havaittavissa suurempaa vaihtelua, mikä saattaa olla seurausta naaras ja koirasnäytteiden eroavasta määrästä.

a)

b)



Kuva 6 - Heterotsygotia -histogrammit kullekin taulukossa 1. oleville populaatiolle. a) Hanko ja b) Konnevesi. Vasemmalla esitetään havaittu ja oikealla odotettu heterotsygotia.

3.2.3. F-statistiikka

F-statistiikat laskettiin Konneveden ja Hangon populaatioiden välillä ja lisäksi siten, että populaatiot eroteltiin naaras- ja koirasnäytteisiin (Taulukko 6). F_{IT} ja F_{IS} saivat pääasiassa negatiivisia arvoja, mikä viittaa siihen, että kiiltomato populaatioissa on vähän, jos lainkaan sukusiittoisuutta. Yhdistetyssä aineistossa ja naaraspopulaatiossa oli matala F_{ST}, mikä viittaa vähäiseen erilaistumiseen ja vastaavasti koiraspopulaatiossa erilaistumista ei ollut lainkaan,

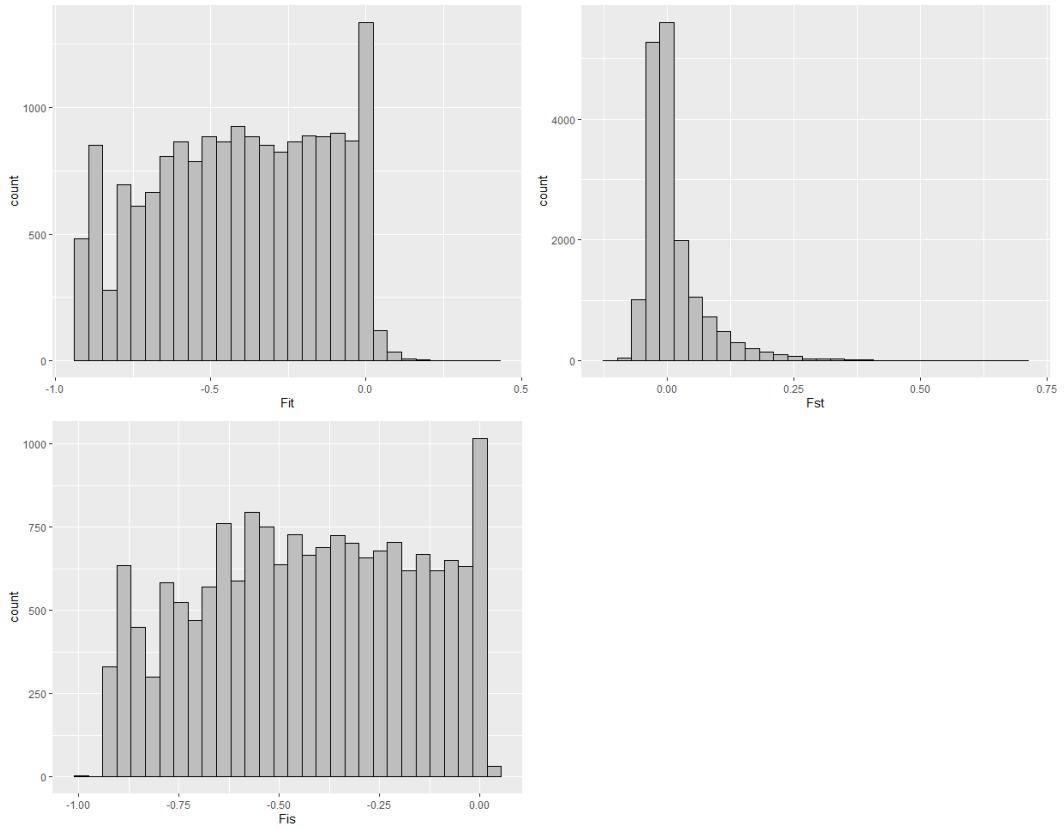
johtuen mahdollisesti näytteiden määrästä, mihin saattoi vaikuttaa koiraiden vähäinen näytemäärä Konneveden populaatiossa.

Taulukko 6 - Keskiarvot kaikkien populaatioiden F-statistiikoista

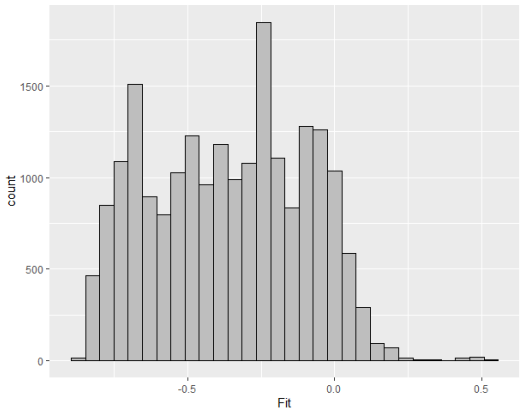
Populaatio	F_{IT}	F_{ST}	F_{IS}
Hanko vs. Konnevesi	-0.4054	0.01214	-0.4238
Hanko vs. Konnevesi - Koiraat	-0.3647	-0.00093454	-0.3790
Hanko vs. Konnevesi - Naaraat	-0.3837	0.01595	-0.4111

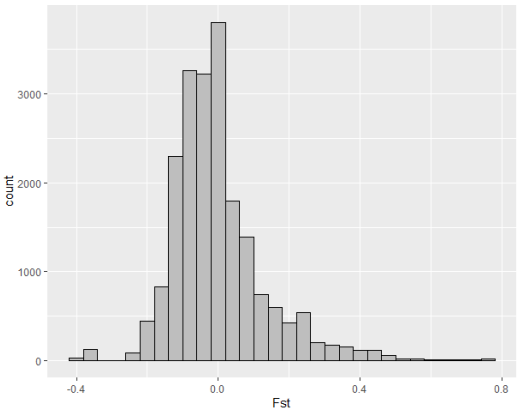
F-statistiikkojen kuvaajista voidaan havaita, että suuri määrä arvoista oli negatiivisia ja pääasiassa vain koiraspopulaatiossa (Kuva 7b) esiintyi positiivisia F_{IT} ja F_{IS} -arvoja, mikä viittaa vähäiseen sukusiittoisuuteen. Naaraspopulaatiossa (Kuva 7c) esiintyi joitain lokuksia, jotka saivat positiivisia F_{IT} -arvoja.

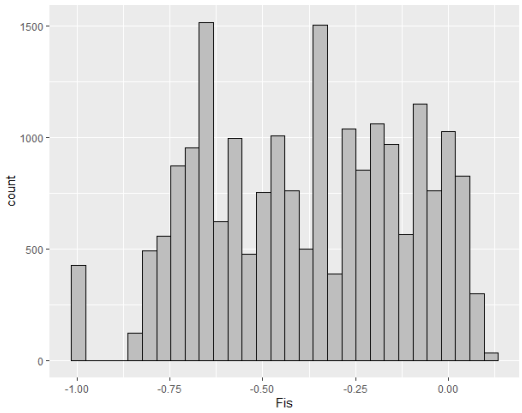
a)



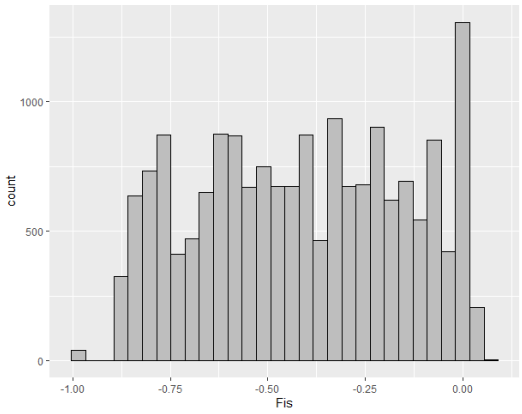
b)

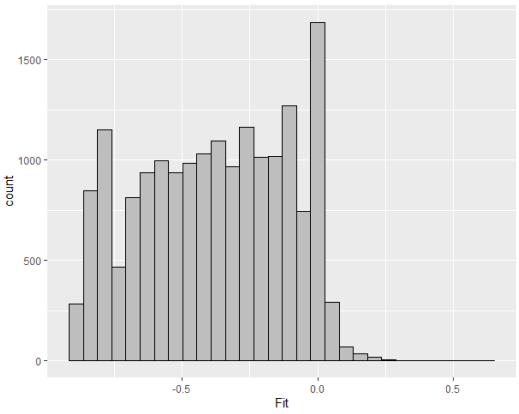


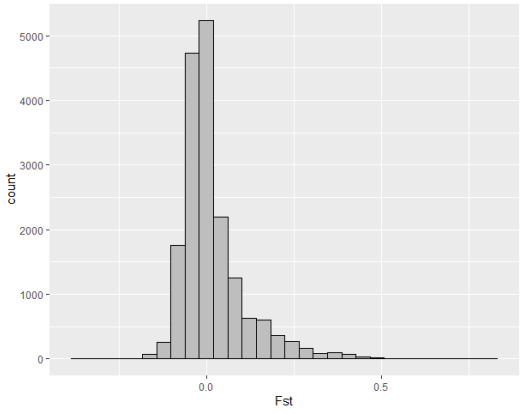




c)







Kuva 7 - Kaikkien taulukosssa 3. olevien populaatioiden F-statistiikkojen kuvaajat. a) Yhdistetty aineisto, b) Koiraat ja c) Naaraat.

3.3. McKinneyn menetelmän suodatus

3.3.1. Suodatus

McKinney-suodatuksen jälkeen yhdistetyn aineiston näytteissä, jotka suodatettiin heterotsygotian suhteen, SNP:n määrä per näyte oli 4322. Vastaavasti lukusuhdepoikkeaman suhteen suodatetuissa näytteissä SNP:n määrä per näyte oli 6365. Näytteissä, jotka suodatettiin molemmilla, oli 4164 SNP:tä per näyte. Konneveden populaatiossa McKinney-suodatuksen jälkeen oli 4519 SNP:tä per näyte ja vastaavasti Hangon populaatiossa suodatuksen jälkeen oli 4453 SNP:tä per näyte. Yhdistetyn aineiston naarasnäytteissä McKinney-suodatuksen jälkeen oli 4340 SNP:tä per näyte ja koirasnäytteissä 5227 SNP:tä per näyte. Suodatus lukusuhdepoikkeaman perusteella ei vaikuttanut itsessään tuloksiin kovinkaan paljon, mutta heterotsygotialla suodattamisella oli odotettu vaikutus tuloksiin ottaen huomioon, että ennen suodattamista havaittu heterotsygotia sai korkeita arvoja yli 0.6.

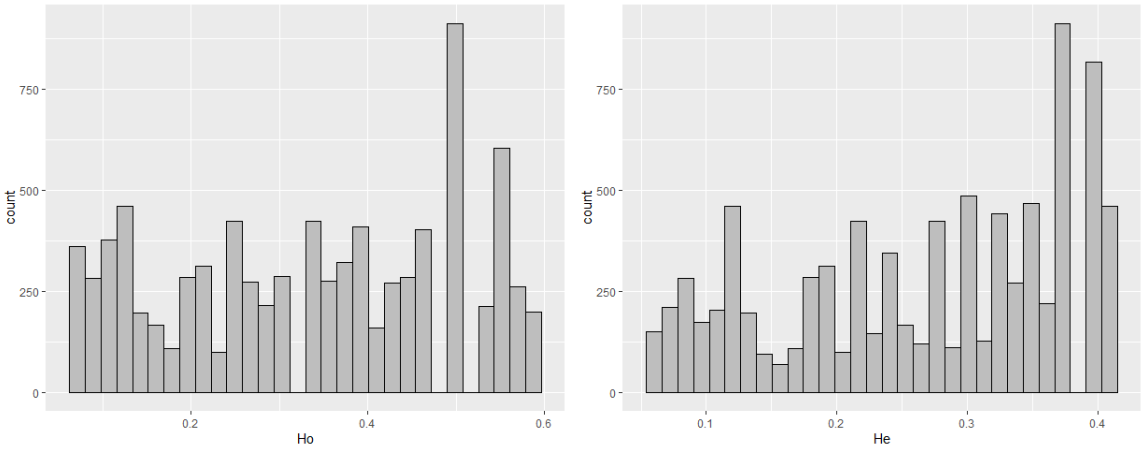
3.3.2. Heterotsygotia

McKinney-suodatettujen näytteiden odotettu heterotsygotia vaihteli 0.26–0.27 välillä ja havaittu heterotsygotia 0.33–0.34 välillä johtuen McKinneyn menetelmän kriteereistä (Taulukko 7). McKinney-suodatus laski saatuja heterotygotian arvoja ennen suodatusta saaduista arvoista noin 0.2 havaitun heterotsygotian suhteen ja 0.1 odotetun heterotsygotian suhteen. Molempien populaatioiden matala heterotsygotia viittaa siihen, että populaatioiden muuntelevat paikat ennen suodatusta olivat pääosin paralogisia ja eivätkä siten oikeasti ole muuntelevia. Aikaisempaan tulokseen verrattaessa Hangon populaatiolla oli alhaisempi odotettu ja havaittu heterotsygotia kuin Konneveden populaatiossa. Heterotsygotian kuvaajat osoittavat paikoin korkeaan heterotsygotiaan varsinkin Konneveden populaatiossa sekä odotetun, että havaitun heterotsygotian suhteen (Kuva 8b).

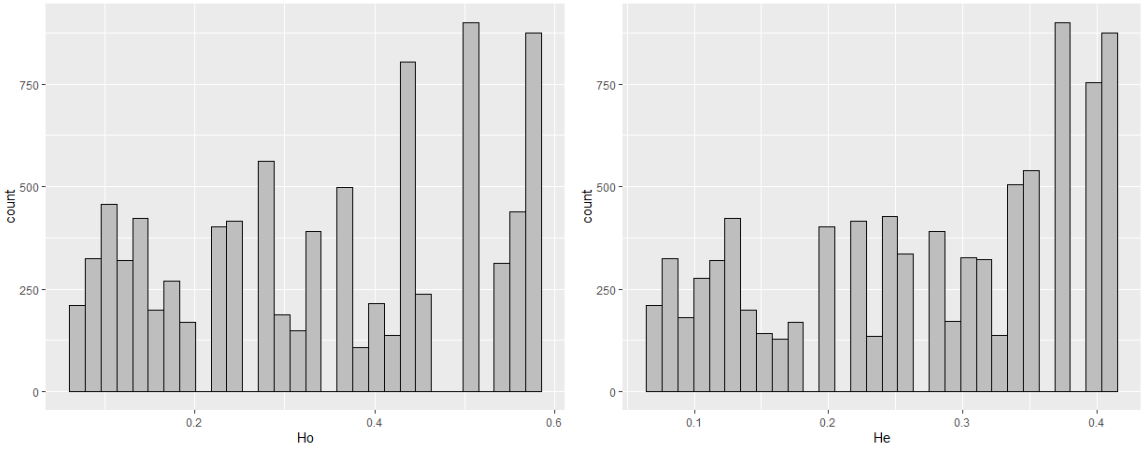
Taulukko 7 - Keskiarvot kaikkien populaatioiden heterotsygotioista McKinney-filteröinnin jälkeen

Populaatio	Ho	He
Hanko	0.3361	0.2670
Konnevesi	0.3425	0.2707

a)



b)



Kuva 8- Heterotsygotia -histogrammit McKinney-suodatetuille populaatiolle. a) Hanko ja b) Konnevesi. Vasemmalla esitetään havaittu ja oikealla odotettu heterotsygotia.

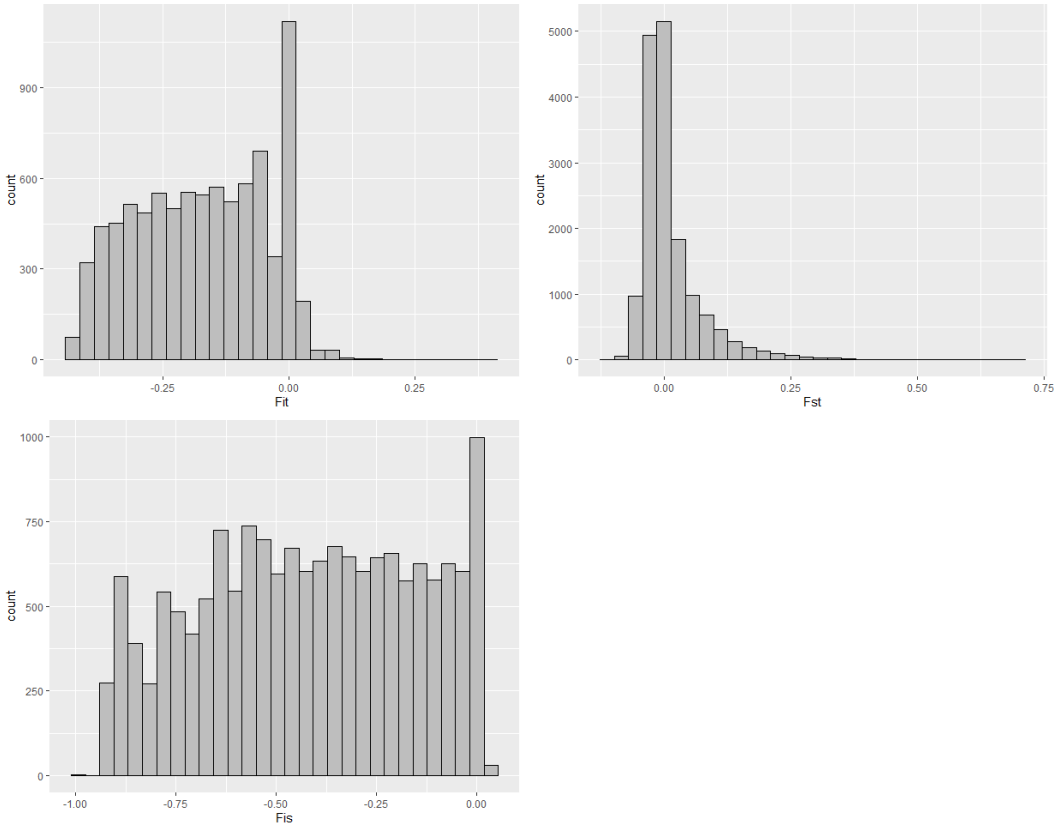
3.3.3. F-Statistiikka

McKinney-suodatetuille Hangon ja Konneveden populaatioille ja niistä erotelluille naaras- ja koiraspopulaatioille laskettiin F-statistiikat, joista F_{ST} sai samanlaisia arvoja kuin ennen suodatusta (Taulukko 8). Suodattaminen havaitulla heterotsygotialla rajoittaa F_{IT} ja F_{IS} arvoja suuresti osoittaen, että populaatioissa oli mahdollisesti ylimäärä paralogeja. Verrattaessa aiempiin tuloksiin heterotsygotian pohjalta suodatetut populaatioiden yksilöt olivat geneettisesti läheisempiä, mutta muutoin tulokset eivät eronneet kovinkaan paljon aikaisemmasta edes histogrammeissa (Kuva 9).

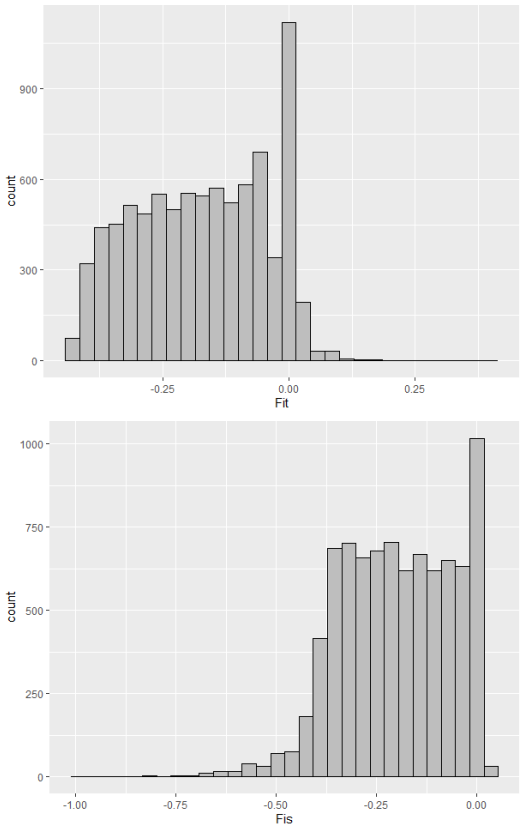
Taulukko 8 - Keskiarvot kaikkien populaatioiden F-statistiikoista McKinney-suodatuksen jälkeen

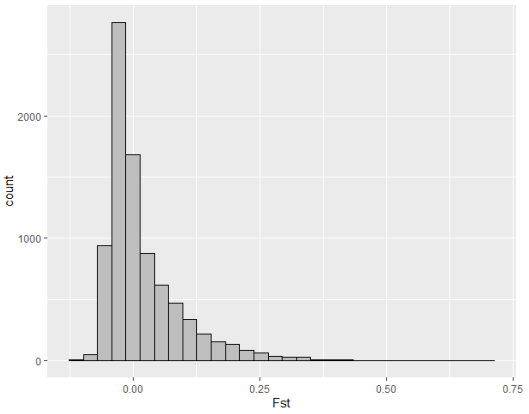
Hanko vs. Konnevesi	F_{IT}	F_{ST}	F_{IS}
H < 0.6	-0.1715	0.01855	-0.1984
-20 < D < 20	-0.4008	0.01201	-0.4190
H < 0.6 & -20 < D < 20	-0.1690	0.01826	-0.1955
Koiraat	-0.1533	-0.002388	-0.1720
Naaraat	-0.1563	0.0233381	-0.1950

a)



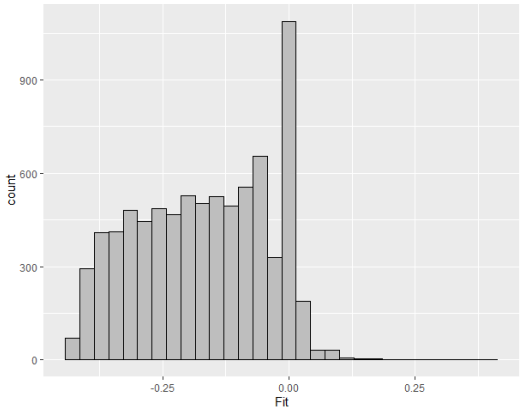
b)

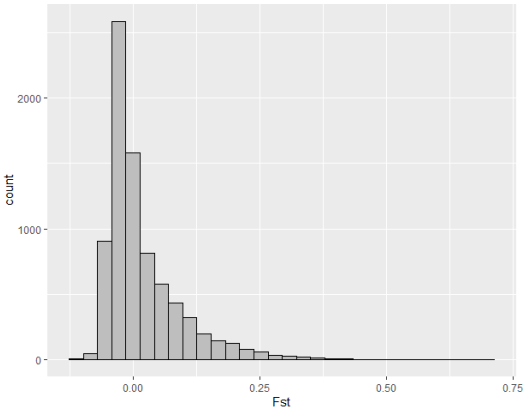

count
900
600
300
0
-0.25
0.00
0.25
Fit

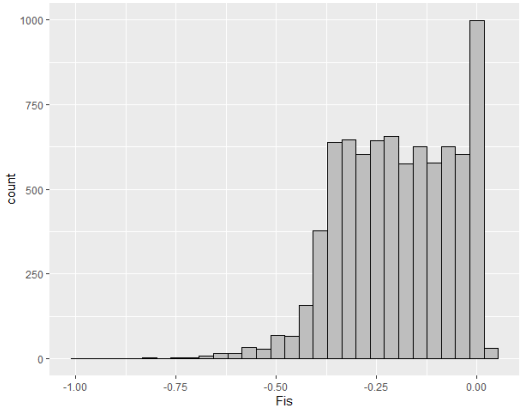

count
2000
1000
0
0.00
0.25
0.50
0.75
Fst

count
1000
750
500
250
0
-1.00
-0.75
-0.50
-0.25
0.00
Fis

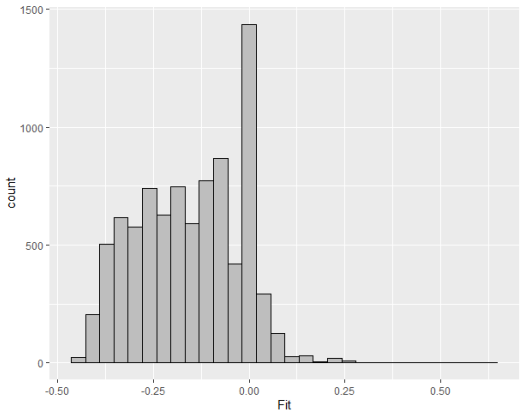
c)

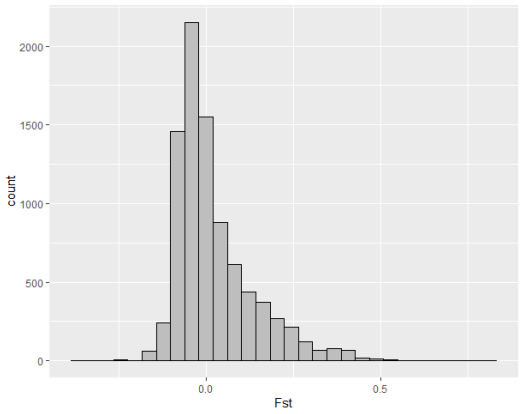

count
900
600
300
0
-0.25
0.00
0.25
Fit

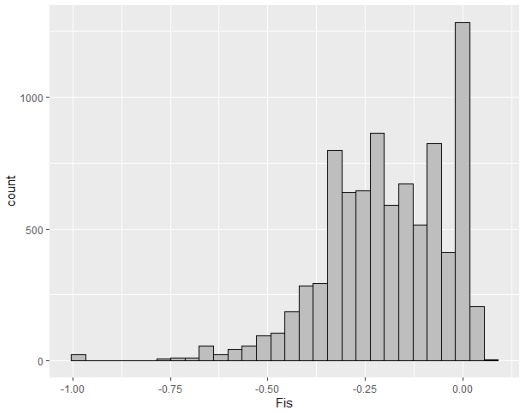

count
2000
1000
0
0.00
0.25
0.50
0.75
Fst


count
1000
750
500
250
0
-1.00
-0.75
-0.50
-0.25
0.00
Fis

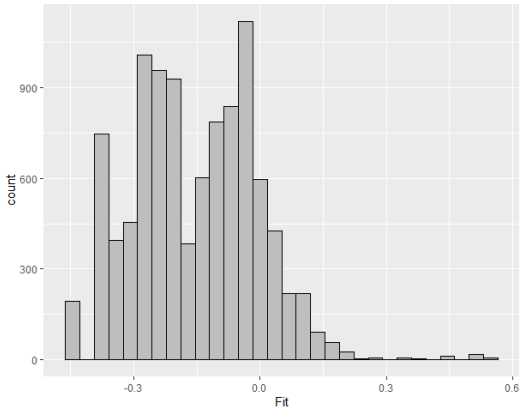
d)

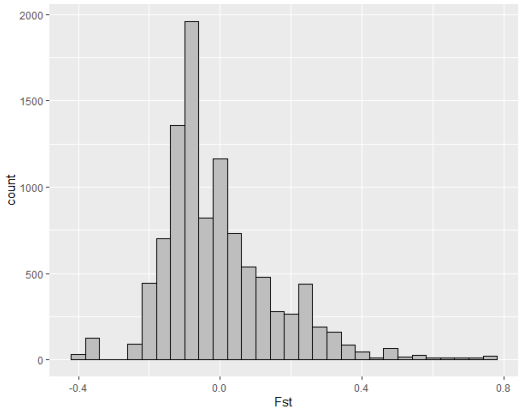


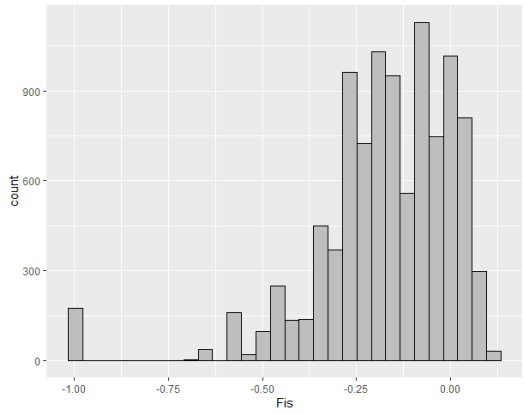




e)







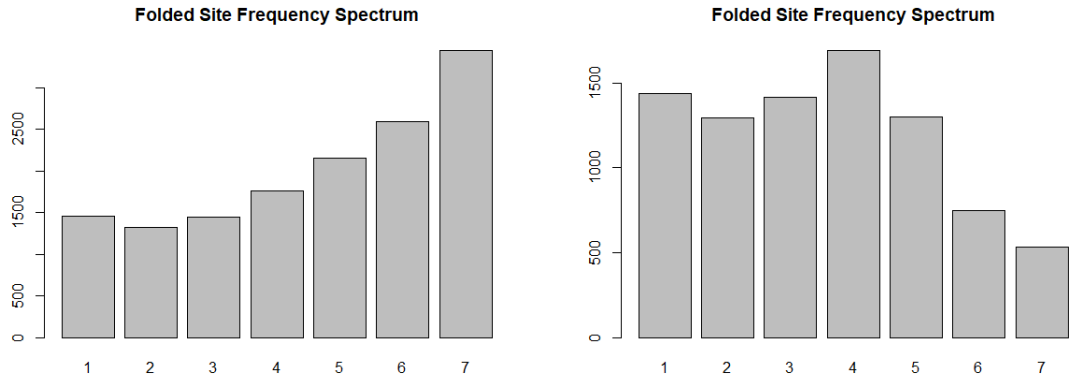
Kuva 9- Kaikkien taulukossa 4. olevien populaatioiden F-statistiikkojen kuvaajat. Yhdistetyn aineiston suodatukset: a) H < 0.6, b) -20 < D < 20 ja c) H < 0.6 & -20 < D < 20. Muut suodatukset: d) Naaraat ja e) Koiraat.

3.4. Alleelifrekvenssispektri

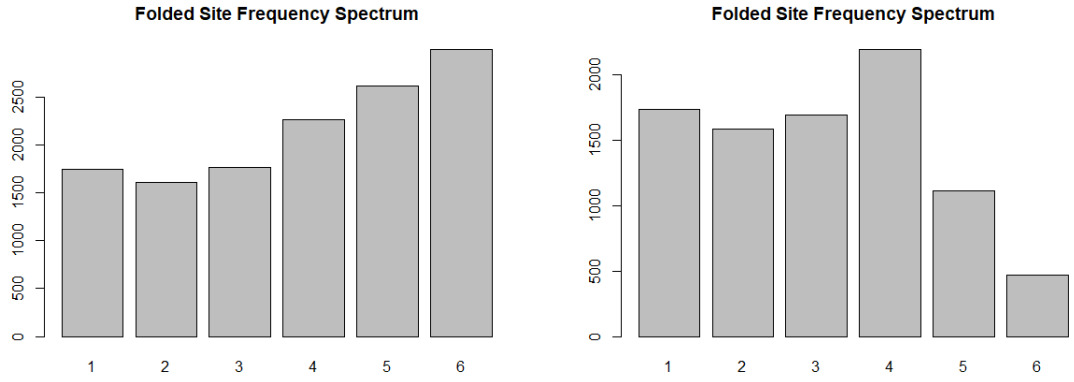
Konneveden ja Hangon näytteissä oli havaittavissa havaitun heterotsygotian vaikutuksesta ennen suodattamista alhainen harvinaisten alleelien määrä, mikä muuttui suodatuksesta tyypillisempään spektrin muotoon eli suurempaan määrään singletoneja verrattaessa muihin frekvensseihin. Kuitenkin frekvenssiluokan 4 kohdalla oli havaittavissa piikki alleelifrekvenssissä kummassakin populaatiossa (Kuva 10), mikä tarkoittaa, että neljässä kromosomissa on paljon harvinaisia alleeleja kummassakin populaatiossa. Konneveden populaatiossa havaittiin 6 kromosomipaikkaa, mikä poikkesi Hangon populaation 7 kromosomipaikasta, mikä saattaa olla seurausta eri määrästä näytteitä ottaen huomioon Konneveden 13 näytettä ja Hangon 15 näytettä.

Ei-suodatettu McKinney

a)



b)



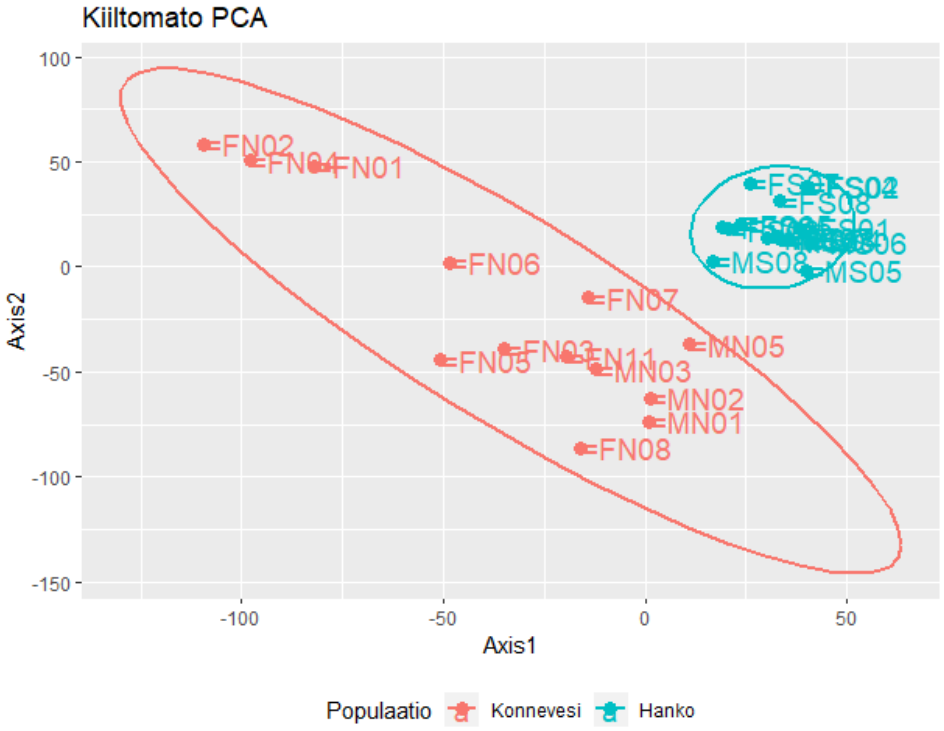
Kuva 10 - a) Hangon ja b) Konneveden ei-suodatettu ja McKinney-suodatettu alleelifrekvenssispektri. X-akselilla kuvataan, kuinka monessa kromosomissa harvinainen alleeli esiintyy. Y-akselilla esitetään harvinaisten alleelien frekvenssi.

3.5. Pääkomponenttianalyysi

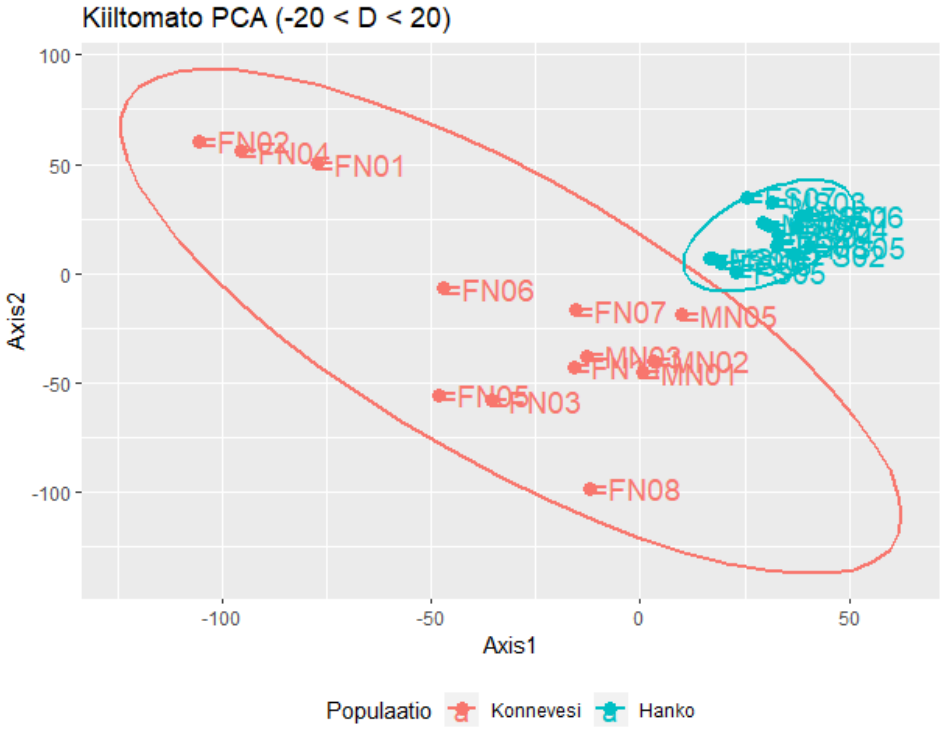
Pääkomponenttianalyysi osoittaa suoraan, että Konneveden ja Hangon populaatiot ovat toisistaan geneettisesti eroavia ryhmiä mikä korreloi aikaisemman tutkimustiedon kanssa ottaen huomioon koiraiden huonon lentokyvyn ja naaraiden vähäisen liikkumisen. Kuitenkin Konneveden populaatiossa oli verrattaessa Hangon populaatioon enemmän muuntelua, mikä saattaa johtua näytteenkeruu alueen koosta Konneveden tutkimusaseman ja Siikakosken rannan välillä. McKinneyn-suodatuksen vaikutuksessa lokusten määrän vähentyessä populaatiot lähestyivät toisiaan siten, että Konneveden yksilöt FN07 ja MN05 olivat geneettisesti

samankaltaisia Hangon populaatioiden kanssa. On kuitenkin huomioitava, että McKinneyn-suodatuksen seurauksena SNPn määrä laskee, mikä vähentää analyysien voimaa.

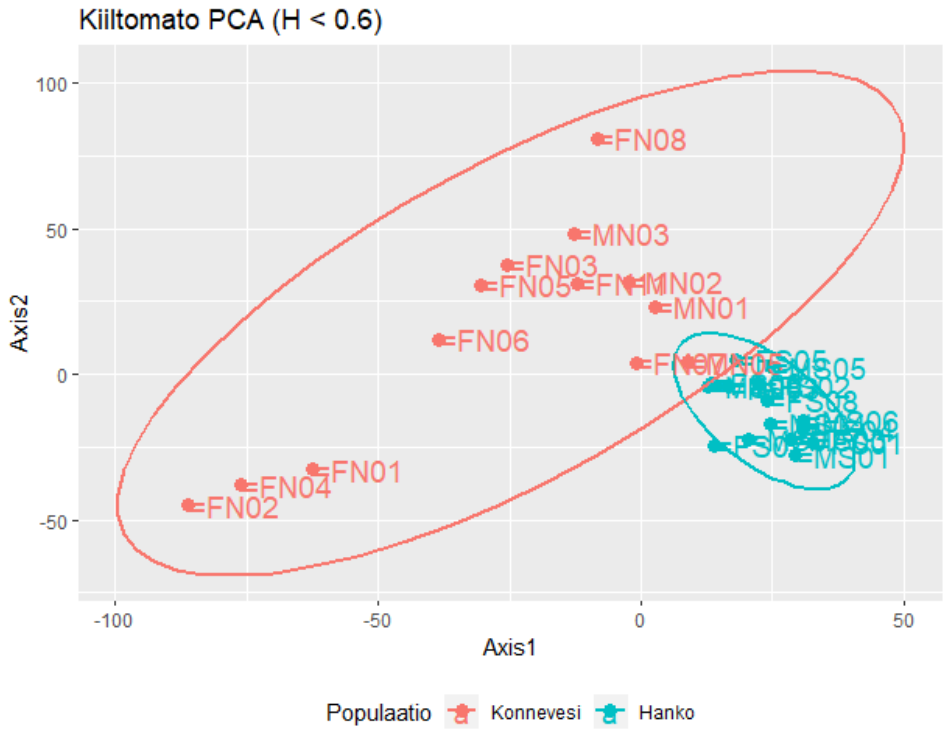
a)



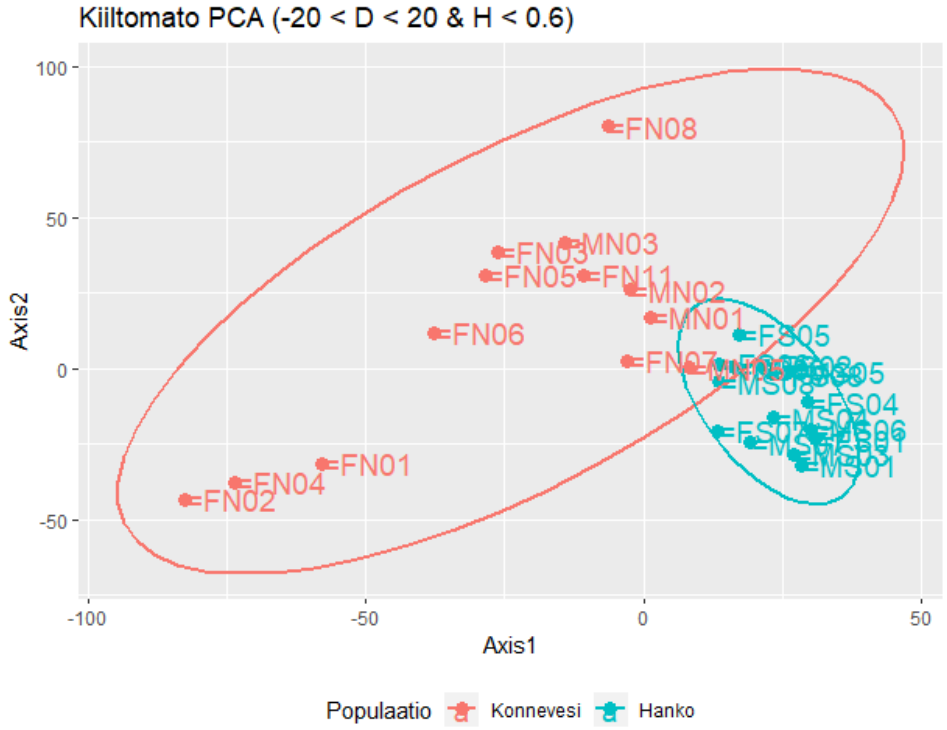
b)



c)



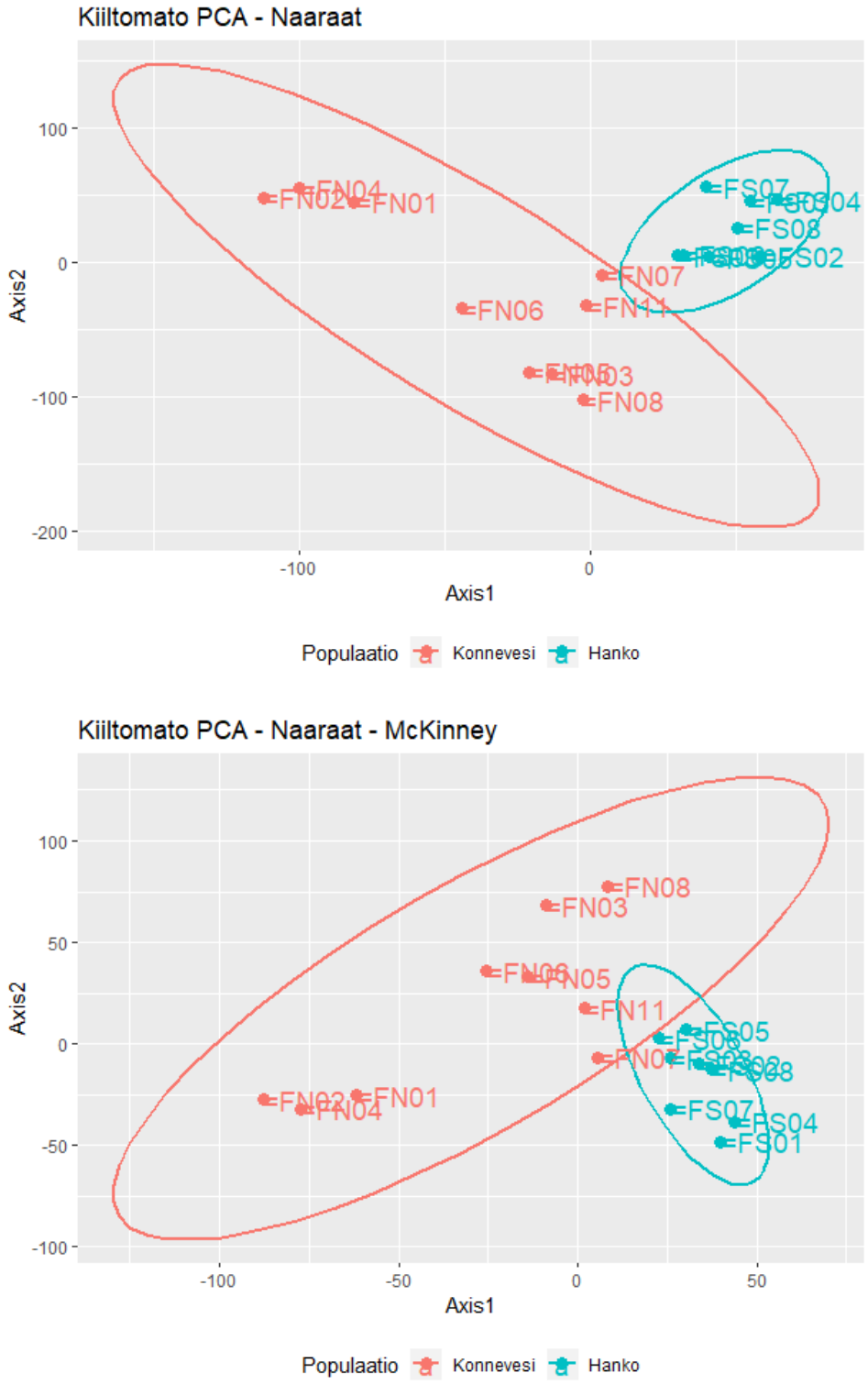
d)



Kuva 11- Yhdistetyn transkriptomin pääkomponenttianalyysin kuvaajat. a) Ei-filteröity, b) -20 < D < 20, c) H < 0.6 ja d) -20 < D < 20 & H < 0.6

Naaraspopulaatioiden pääkomponenttianalyysistä etäisyydet olivat pysyneet samankaltaisina kokonaispopulaatioiden kanssa. Kuitenkin siitä voidaan huomata, että Konneveden naaraista

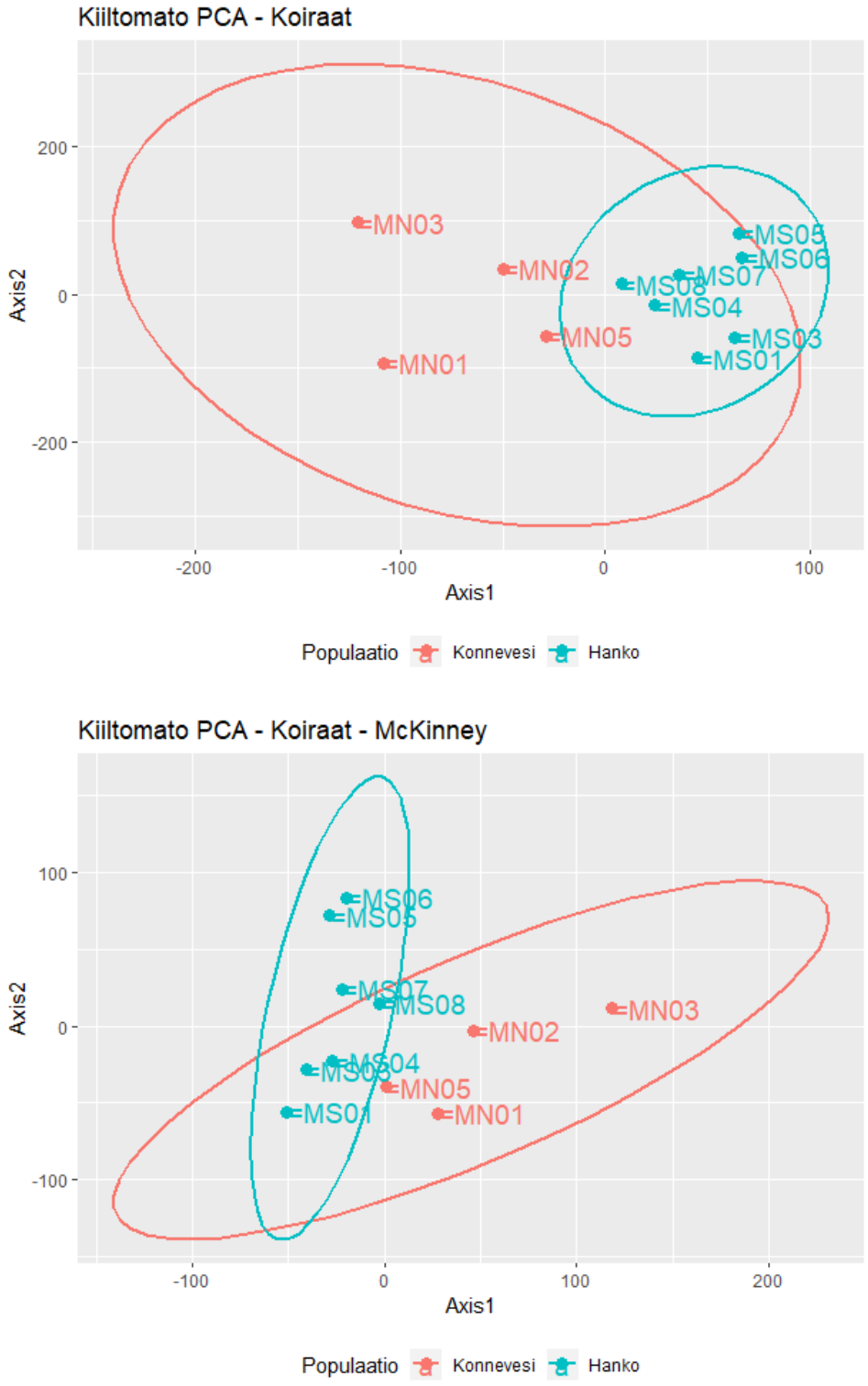
yksilöt FN01, FN02 ja FN04 muodostivat yhden ryhmän, FN03, FN05 ja FN08 muodostivat ryhmän sekä FN07 ja FN11 muodostivat, mikä saattaa olla seurausta sukulaisuudesta. Kuitenkin McKinney-suodatuksen seurauksena FN03, FN05 ja FN08 muodostama ryhmittymä hajosi, mutta FN05 ja FN06 olivat vierekkäin.



Kuva 12 - Yhteispopulaation naaraiden ei-suodatetun ja McKinney-suodatetun pääkomponenttianalyysin kuvaajat.

Populaatioiden koiraiden pääkomponenttianalyysissä Konneveden yksilöt olivat etäisiä toisiinsa nähden, mutta niiden muuntelu vaikutti olevan samankaltainen Hangon yksilöiden

kanssa. Koiraat jakautuivat oman populaationsa sisällä ryhmiin Hangossa, jossa MS05-06, MS04 ja MS07-08 sekä MS01 ja MS03 muodostavat ryhmän, kun taas Konnevedellä selvää ryhmittymistä ei ollut. McKinney-suodatuksella ryhmittyminen muuttui. MS04 oli MS03:n ja MS01 ryhmittymässä ja Konnevedellä MN05 ja MN01 lähestyivät toisiaan.



Kuva 13 - Yhteispopulaation koiraiden ei-suodatetun ja McKinney-suodatetun pääkomponenttianalyysin kuvaajat.

4. Pohdinta

4.1. Suodatuksen vaikutus transkriptomidataan

Tämän tutkielman päätavoitteena oli selvittää, kuinka transkriptomidata soveltuu kiiltomatojen (*L. noctiluca*) populaatiogeneettiseen analyysiin ja samalla selvittää, miten suodatus vaikuttaa aineiston laatuun.

Päällekkäisyyden poistamisen menetelmällä CD-HIT oli selvä vaikutus tulosten laatuun ja ennen CD-HIT-käsittelyä aineiston Kallisto-analyysissä näkynyt poikkeama, joka havaittiin ExN50 ekspressoitujen E10 – E20 ja E5 – E15 frekvenssien välillä (Kuvat 1 ja 2), muuttui aineistosta käsittelyn jälkeen (Kuvat 3 ja 4). Hangossa poikkeaman contigin pituus puolittui ja molemmissa populaatioissa expressoitu alue muuttui Hangossa välille E15 – E23 ja Konnevedellä E5 – E9. Tämä viittaa siihen, että tutkimusaineistossa esiintyy päällekkäisyyttä vähän ekspressoiduissa contigeissa. Annotaatiosta saatiin selville, että kummassakin populaatiossa kasvupiikin ekspression alueella esiintyvät contigit vastaavat vitellogeniiniin, joka liittyy munasolun muodostukseen. Korkea ekspressio yhdessä vähäisen sekvenssimuuntelun kanssa voisi selittää sen, miksi tämän geenin transkriptista on saatu koostettua iso contig, mikä voisi selittää kasvupiikin. Kuitenkin on tyypillistä havaita RNASeq-tutkimuksissa redundanssia, mikä johtuu menetelmän lyhyiden lukemaparien ja transkriptomi-datan laajuudesta (Góngora-Castillo & Buell, 2013; Wolf 2013).

Käytettäessä McKinneyn-menetelmää suodatuksella oli selvä vaikutus tutkimusaineistoon. Itse suodatus vähensi Konneveden SNP:n määrää noin 47 % ja Hangossa 40 % sekä vastaavasti yhdistetyn aineiston koirailla 49 % ja naarailla 44 %, mikä vähentää tulosten voimaa merkittävästi, mutta parantaa niiden luotettavuutta. Ennen sekvensointia datasta havaittavat populaatioiden eroavaisuudet osoittavat, että Hangon populaatiossa on korkeampi heterotsygotia kuin Konneveden populaatiossa. Kuitenkin suodatuksen jälkeen tilanne muuttui siten, että Konneveden populaation heterotsygotia oli korkeampi. Tämä viittaa siihen, että ennen suodatusta Hangon korkeampi heterotsygotia johtui kummankin populaation mahdollisesta paralogien ylijäämästä. Paralogien suhteen suodattaminen ei vaikuttanut huomattavasti F_{ST}:n arvoihin, mutta F_{IT} ja F_{IS} -arvoihin se vaikutti. Suodattaminen McKinneyn-menetelmällä vähensi kummankin populaation F_{IT}- ja F_{IS}-arvoista noin 75 %, mikä viittaa suureen paralogian ylijäämään ennen suodatusta.

Molempien populaatioiden alleelifrekvenssispektreissä esiintyi vääristymää perussuodatuksen jälkeen, mikä saattaa olla seurausta heterotsygotian ylijäämästä. McKinneyn-menetelmällä

suodatus muutti alleelifrekvenssispektrin kuvaajan vastaamaan spektrille odotettua harvinaistenalleelien frekvenssiä (Kuva 10), mikä vahvasti viittaa siihen, että kummassakin populaatiossa korkea heterotsygotia johtui paralogien ylijäämästä. Merkittävä paralogien määrä viittaa paralogisien geenien läsnäoloon kiiltomadoissa, minkä lisätutkiminen on mielestäni tarpeellista. Vastaavaa vääristymää alleelifrekvenssispektreissä esiintyi myös sembramännyillä johtuen suuresta paralogien määrässä, joka viittaa paralogi-ongelman yleisyyteen transkriptomitutkimuksissa (Rellstab ym. 2018).

Pääkomponenttianalyyseistä huomattiin, että suodatuksen vaikutus populaatiorakenteeseen ei ole kovinkaan suurta. Populaatiot lähestyvät toisiaan McKinneyn-menetelmällä suodattamisen vaikutuksesta, mutta itse suodatuksella ei lukusuhdepoikkeamalla tai heterotsygotialla näytä olevan suoraa vaikutusta kuvaajiin. Kuitenkin ottaen huomioon SNP:n vähentymisen suodatuksesta on todennäköistä, että pienempi aineisto määrä vaikuttaa tuloksiin. Suodatuksen vaikutus on merkittävintä populaatioiden koiraiden välillä, mikä luultavasti johtuu jo valmiiksi pienemmästä näytemäärästä ja SNP:n määrän vähentymisessä populaatiossa, kun otetaan huomioon, kuinka monta SNP:ä suodattui pois aineistosta.

4.2. Populaatiorakenne

Toinen tavoite tässä tutkimuksessa oli selvittää, millaista geneettistä erilaistumista Hangon ja Konneveden kiiltomatopopulaatioiden välillä on. Oletuksena oli, että populaatioiden välillä on erilaistumista, koska naaraat liikkuvat vähän kuoriutumisen jälkeen ja koiraat eivät ole hyviä lentäjiä (Tyler 2002). Erilaistumista selvitettiin määrittämällä heterotsygotiat kummastakin populaatiosta ja niiden väliset F-statistiikat. Tässä huomattiin esiintyvän odottamattoman korkeaa heterotsygotiaa, mikä on mahdollisesti seurausta RNASeq-menetelmälle tyypillisestä paralogien ylijäämästä.

Kummankin populaation heterotsygotian arvot ovat keskenään samankaltaisia ennen suodatusta ja sen jälkeen. Havaitun ja odotetun heterotsygotian arvot ovat hieman korkeampia Konnevedellä kuin Hangossa, mutta ei merkittävästi. Ottaen huomioon populaatioiden korkean heterozygoottien määrän on oletettavaa, että populaatioissa on vähän sukusiittoisuutta.

Populaatioiden F-statistiikat saivat negatiivisia F_{IS} ja F_{IT} arvoja, mikä viittaa siihen, että populaatioissa ei ole sukusiittoisuutta vaan yksilöt eroavat toisistaan. Tämä johtuu mahdollisesti populaatioiden etäisyydestä toisiinsa. Syynä tähän on kiiltomatojen vähäinen liikkuvuus aikuisvaiheessa ja migraation tapahtuminen toukkavaiheessa (Tyler 2002; Hickmott & Tyler, 2011) sekä mahdollinen ihmisen tekemän keinotekoisen valon vaikutus niiden elinympäristöihin

(Ineichen & Rüttiman, 2012; Elgert ym. 2020), mikä rajoittaa kiiltomatojen migraatiota vähentämällä niiden kykyä lisääntyä. Populaatioiden F_{ST}-arvot olivat yllättävän matalia ottaen huomioon populaatioiden eristyneisyyden. Populaatioiden välillä on vähän geneettistä erilaistumista, mikä mahdollisesti johtuu siitä, että kiiltomadon toukkavaiheen migraatio estää erilaistumisen populaatioiden välillä, mutta tämä vaatii lisätutkimusta. Populaatioiden koiraiden välillä ei havaittu minkäänlaista geneettistä erilaistumista, mikä luultavasti johtuu suuresta näytekoon erosta koirasnäytteissä (Konnevedellä oli vain neljä koirasta Hangon seitsemään).

Alleelifrekvenssispektreissä esiintyy kummassakin populaatiossa samankaltainen piikki saman alleeliluokan (4) kohdalla, mutta Konnevedellä frekvenssejä on suuremman näytekoon takia enemmän. Frekvenssin nousu on mahdollisesti seurausta pullonkaulan ja populaation kasvun yhteisvaikutuksesta, mikä kertoo kiiltomatojen metapopulaation (Hanski 1998) kaltaisesta elinkierrosta ja populaatioiden migraatiosta (Tyler 2002).

Hangon ja Konneveden populaatiot erottautuvat toisistaan pääkomponenttianalyysin kuvaajissa selvästi, mikä vastaa alkuperäistä huomiota kiiltomatojen liikkuvuudesta ja näytteenottopaikkojen välisestä etäisyydestä. Vastaavasti Konnevedeltä kerätyissä yksilöissä oli enemmän populaation sisäistä muuntelua, mikä johtuu todennäköisesti Konneveden näytteenottopaikkojen välisestä etäisyydestä. Naaraiden väliset eroavaisuudet vastaavat kokonaispopulaatiota, mutta koirailla näkyy selvää samankaltaisuutta populaatioissa, mikä vastaa F-statistiikan tuloksia.

5. Yhteenveto

Tämän tutkimuksen päätavoitteena oli tarkastella, kuinka transkriptomi-aineisto soveltuu populaatiogeneettiseen analyysiin ja miten suodatus vaikuttaa aineiston laatuun. Sen lisäksi tämän tutkimuksen toisena päätavoitteena oli selvittää, minkälaista geneettistä muuntelua Konneveden ja Hangon populaatioiden välillä on.

Transkriptomi-aineiston soveltuvuutta tutkittiin Viljakainen ym. 2020 tutkimuksessa kerätyn Hangon ja Konneveden *Lampyris noctiluca* - populaatiota koskevan aineiston avulla. Laaduntarkistus ja siihen liittyvät analyysit oli tehty aikaisemmin. Käytössä olevan transkriptomi-aineiston kokonaisuus arvioitiin ja sen jälkeen yksilöiden sekvenssit linjattiin yhdistettyyn aineistoon ja tehtiin suodatukset. Tämän suodatuksen lisäksi aineistolle tehtiin McKinneyn-menetelmällä paralogien suodatus (McKinney ym. 2017). Ennen McKinneyn-menetelmän suodatusta tuloksissa oli korkeita heterotsygotian arvoja sekä voimakkaan negatiivisia F_{IS}- ja F_{IT}-statistiikkojen arvoja, jotka laskivat merkitsevästi suodatuksen jälkeen, joten niiden pääteltiin olevan

seurausta ylijäämästä paralogeja. Samoin alleelifrekvensseissä havaittiin vääristymää ennen suodatusta, mikä osoittaa, että kiiltomadoilla on paralogisia geenejä. Koska paralogian määrä on aineistossa suuri, niin kiiltomadon paralogia kaipaa lisätutkimusta.

Konneveden ja Hangon populaatioiden välistä geneettistä muuntelua tutkittiin heterotsygotialla ja F-statistiikoilla. Kummankin populaation heterotsygotia sai lähes samanlaiset arvot ennen suodatusta ja sen jälkeen. Konneveden heterotsygotia oli vain vähän suurempi kuin Hangossa.

F-statistiikat olivat populaatioissa yllättävän matalat, vaikka kiiltomatopopulaatioiden välinen etäisyys on suuri ja kiiltomadolla on vain vähän migraatiota. Hangon ja Konneveden populaatiot eroavat toisistaan enemmän kuin on tyypillistä satunnaisesti pariutuvissa populaatioissa. Populaatioiden välillä on vähän geneettistä muuntelua ja populaatioiden koiraiden välillä ei ole muuntelua. Pääkomponenttianalyysissä populaatiot erottuivat toisistaan selvästi, mutta koirailla eroja oli vähemmän kuin naarailla. Populaatioiden välisen muuntelun määrän vähäisyys ja koiraissa esiintyvän muuntelun puute saattavat johtua siitä, että varsinkin koiraista otettujen näytteiden määrä erosi suuresti. Konneveden populaatiosta oli 4 näytettä ja vastaavasti Hangossa 7. Populaatioissa esiintyy tämän tutkimusten tulosten perusteella yllättävän vähän muuntelua, joten lisätutkimus saattaa olla mielekästä.

6. Kiitokset

Pääasiallisesti haluan kiittää ohjaajiani Lumi Viljakaista ja Tanja Pyhäjärveä tämän työn ohjaamisesta. Erityinen kiitos Lumi Viljakaiselle hyvästä ohjauksessa tutkielman kirjoittamisessa, sen taustatutkimuksissa ja analyyseissä esiintyneiden ongelmien kanssa sekä tämän tutkielman suuntausta koskevista neuvoista. Lisäksi Tanja Pyhäjärvelle kiitokset avusta bioinformaattisten analyysien ja ohjelmien käytössä esiintyneiden ongelmien ratkaisemisessa sekä tulosten ja menetelmien selkeyttämisessä. Kiitokset myös aineiston keräämisestä Anna-Maria Borshagovskille ja Sami Saarenpäälle. Erittäin suuret kiitokset myös professori Arja Kaitalalle ja kiiltomatojen tutkimusryhmälle tutkimuskirjallisuudesta.

7. Kirjallisuus

1. Allendorf, F. W., & Leary, R. F. (1986). Heterozygosity and fitness in natural populations of animals. *Conservation biology: the science of scarcity and diversity*, *57*, 58–72.
2. Attard, C. R., Beheregaray, L. B., Sandoval-Castillo, J., Jenner, K. C. S., Gill, P. C., Jenner, M. N. M., ... & Möller, L. M. (2018). From conservation genetics to conservation genomics: a genome-wide assessment of blue whales (Balaenoptera musculus) in Australian feeding aggregations. *Royal Society open science*, *5*(1), 170925.
3. Benestan, L. M., Ferchaud, A. L., Hohenlohe, P. A., Garner, B. A., Naylor, G. J., Baums, I. B., ... & Luikart, G. (2016). Conservation genomics of natural and managed populations: building a conceptual and practical framework. *Molecular ecology*, *25*(13), 2967–2977.
4. Bohonak, A. J. (1999). Dispersal, gene flow, and population structure. *The Quarterly review of biology*, *74*(1), 21–45.
5. Borshagovski, A. M., Baudry, G., Hopkins, J., & Kaitala, A. (2019). Pale by comparison: competitive interactions between signaling female glow-worms. *Behavioral Ecology*, *30*(1), 20-26.
6. Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, *34*(5), 525–527.
7. Bustamante, C. D., Wakeley, J., Sawyer, S., & Hartl, D. L. (2001). Directional selection and the site-frequency spectrum. *Genetics*, *159*(4), 1779–1788.
8. Chen, B., Cole, J. W., & Grond-Ginsbach, C. (2017). Departure from Hardy Weinberg equilibrium and genotyping error. *Frontiers in genetics*, *8*, 167.
9. Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PloS one*, *12*(12), e0190152.
10. Cutter, A. D. (2019). A primer of molecular population genetics. Oxford University Press.
11. Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... & McVean, G. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156-2158.
12. Day, J. C., Chaichi, M. J., Najafil, I., & Whiteley, A. S. (2006). Genomic structure of the luciferase gene from the bioluminescent beetle, Nyctophila cf. caucasica. *Journal of Insect Science*, *6*(1).

13. De Cock, R., & Matthysen, E. (2003). Glow-worm larvae bioluminescence (Coleoptera: Lampyridae) operates as an aposematic signal upon toads (Bufo bufo). *Behavioral Ecology*, *14*(1), 103–108.
14. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15-21.
15. Dray, S., & Dufour, A. B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of statistical software*, *22*(4), 1–20.
16. Elgert, C., Hopkins, J., Kaitala, A., & Candolin, U. (2020). Reproduction under light pollution: maladaptive response to spatial variation in artificial light in a glow-worm. *Proceedings of the Royal Society B*, *287*(1931), 20200806.
17. Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in ecology & evolution*, *29*(1), 51–63.
18. Frankham, R., Ballou, S. E. J. D., Briscoe, D. A., & Ballou, J. D. (2002). *Introduction to conservation genetics*. Cambridge university press.
19. Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150-3152.
20. Gayral, P., Melo-Ferreira, J., Glemin, S., Bierne, N., Carneiro, M., Nabholz, B., ... & Belkhir, K. (2013). Reference-free population genomics from next-generation transcriptome data and the vertebrate–invertebrate gap. *PLoS Genet, 9*(4), e1003457.
21. Gopalakrishnan, S., Castruita, J. A. S., Sinding, M. H. S., Kuderna, L. F., Räikkönen, J., Petersen, B., ... & Hansen, A. J. (2017). The wolf reference genome sequence (Canis lupus lupus) and its implications for Canis spp. population genomics. *BMC genomics*, *18*(1), 1–11.
22. Góngora-Castillo, E., & Buell, C. R. (2013). Bioinformatics challenges in de novo transcriptome assembly using short read sequences in the absence of a reference genome sequence. *Natural product reports*, *30*(4), 490–500.
23. Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... & Chen, Z. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature biotechnology*, *29*(7), 644.
24. Graffelman, J., Jain, D., & Weir, B. (2017). A genome-wide study of Hardy–Weinberg equilibrium with next generation sequence data. *Human Genetics*, *136*(6), 727–741.
25. Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., ... & MacManes, M. D. (2013). De novo transcript sequence reconstruction from RNA-

seq using the Trinity platform for reference generation and analysis. *Nature protocols*, *8*(8), 1494–1512.

26. Hanski, I. (1998). Metapopulation dynamics. *Nature*, *396*(6706), 41-49.
27. Hansson, B., & Westerberg, L. (2002). On the correlation between heterozygosity and fitness in natural populations. *Molecular ecology*, *11*(12), 2467–2474.
28. Hickmott, W., & Tyler, J. (2011). Seasonal variation in the female display period of the glow-worm *Lampyris noctiluca* L.(Coleoptera: Lampyridae). *Lampyrid, 1*, 14–21.
29. Hopkins, J., Baudry, G., Candolin, U., & Kaitala, A. (2015). I'm sexy and I glow it: female ornamentation in a nocturnal capital breeder. *Biology letters, 11*(10), 20150599.
30. Horne, J. (2011). Unfit or just unlucky? A sexual difference in the timing of adult emergence in the Glow-worm *Lampyris noctiluca* L.(Coleoptera: Lampyridae). *Lampyrid*, *1*, 9–13
31. Horne, J., Horne, A. & Tyler, J. (2017) The influence of female body weight on egg production in the glow-worm *Lampyris noctiluca* (Coleoptera: Lampyridae). *Lampyrid*, *4*, 36-39.
32. Horne, J., & Horne, A. (2017). Larval development rates in the glow-worm *Lampyris noctiluca* (L.). *Lampyrid*, *4*, 55–58.
33. Hudson, M. E. (2008). Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular ecology resources, 8*(1), 3–17.
34. Hrdlickova, R., Toloue, M., & Tian, B. (2017). RNA-Seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA*, *8*(1), e1364.
35. Ineichen, S., & Rüttimann, B. (2012). Impact of artificial light on the distribution of the common European glow-worm, Lampyris noctiluca (Coleoptera: Lampyridae). *Lampyrid*, *2*, 31–36.
36. Jakobsson, M., Edge, M. D., & Rosenberg, N. A. (2013). The relationship between FST and the frequency of the most frequent allele. *Genetics*, *193*(2), 515–528.
37. Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405.
38. Knaus, B. J., & Grünwald, N. J. (2017). vcfr: a package to manipulate and visualize variant call format data in R. *Molecular ecology resources*, *17*(1), 44-53.
39. Kuraku, S. (2010). Palaeophylogenomics of the vertebrate ancestor—impact of hidden paralogy on hagfish and lamprey gene phylogeny.
40. Lehtonen, T. K., & Kaitala, A. (2020). Leave me alone: solitary females attract more mates in a nocturnal insect. *Behavioral Ecology*.

41. Linck, E., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources, 19*(3), 639-647.
42. Luikart, G., Kardos, M., Hand, B. K., Rajora, O. P., Aitken, S. N., & Hohenlohe, P. A. (2018). Population genomics: advancing understanding of nature. In *Population genomics* (pp. 3-79). Springer, Cham.
43. Ma, S., & Dai, Y. (2011). Principal component analysis based methods in bioinformatics studies. *Briefings in bioinformatics, 12*(6), 714-722.
44. Marth, G. T., Czabarka, E., Murvai, J., & Sherry, S. T. (2004). The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, *166*(1), 351–372.
45. Mastrochirico-Filho, V. A., Hata, M. E., Sato, L. S., Jorge, P. H., Foresti, F., Rodriguez, M. V., ... & Hashimoto, D. T. (2016). SNP discovery from liver transcriptome in the fish Piaractus mesopotamicus. *Conservation genetics resources*, *8*(2), 109–114.
46. McKinney, G. J., Waples, R. K., Pascal, C. E., Seeb, L. W., & Seeb, J. E. (2018). Resolving allele dosage in duplicated loci using genotyping-by-sequencing data: A path forward for population genetic analysis. *Molecular ecology resources*, *18*(3), 570–579.
47. McKinney, G. J., Waples, R. K., Seeb, L. W., & Seeb, J. E. (2017). Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources*, *17*(4), 656–669.
48. Meirmans, P. G., & Hedrick, P. W. (2011). Assessing population structure: FST and related measures. *Molecular ecology resources*, *11*(1), 5–18.
49. Meisner, J., & Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*, *210*(2), 719–731.
50. Oba, Y., Konishi, K., Yano, D., Shibata, H., Kato, D. I., & Shirai, T. (2019). Resurrecting the ancient glow. *BioRxiv*, 778688.
51. Paradis, E. (2010). pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics*, *26*(3), 419–420.
52. Platt, A., Horton, M., Huang, Y. S., Li, Y., Anastasio, A. E., Mulyati, N. W., ... & Dunning, M. (2010). The scale of population structure in Arabidopsis thaliana. *PLoS Genet*, *6*(2), e1000843.

53. Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., ... & Shakir, K. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, 201178.
54. Reich, D., Price, A. L., & Patterson, N. (2008). Principal component analysis of genetic data. *Nature genetics, 40*(5), 491–492.
55. Rellstab, C., Dauphin, B., Zoller, S., Brodbeck, S., & Gugerli, F. (2019). Using transcriptome sequencing and pooled exome capture to study local adaptation in the giga-genome of Pinus cembra. *Molecular ecology resources*, *19*(2), 536–551.
56. Sagegami-Oba, R., Takahashi, N., & Oba, Y. (2007). The evolutionary process of bioluminescence and aposematism in cantharoid beetles (Coleoptera: Elateroidea) inferred by the analysis of 18S ribosomal DNA. *Gene*, *400*(1–2), 104-113.
57. Sarah, G., Homa, F., Pointet, S., Contreras, S., Sabot, F., Nabholz, B., ... & Sauvage, C. (2017). A large set of 26 new reference transcriptomes dedicated to comparative population genomics in crops and wild relatives. Molecular ecology resources, 17(3), 565–580.
58. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210-3212.
59. Smith, M. L., & Hahn, M. W. (2020). New approaches for inferring phylogenies in the presence of paralogs. *Trends in Genetics*.
60. Sun, Y., Li, Q., Wei, H., Wang, G., Chen, J., & Li, P. (2018). Single nucleotide polymorphism identification in growth-related genes from the transcriptome of the fish Ancherythroculter nigrocauda. *Conservation Genetics Resources*, *10*(2), 153–155.
61. Tisi, L. C., De Cock, R., Stewart, A. J., Booth, D., & Day, J. C. (2014). Bioluminescent leakage throughout the body of the glow-worm *Lampyris noctiluca* (Coleoptera: Lampyridae). *Entomologia generalis*, *35*(1), 47–51.
62. Tyler J. 2002. The glow-worm. Kent (UK): Lakeside Printing Ltd.
63. Viljakainen, L., Borshagovski, A. M., Saarenpää, S., Kaitala, A., & Jurvansuu, J. (2020). Identification and characterisation of common glow-worm RNA viruses. *Virus Genes*, 1-13.
64. Wang, K., Hong, W., Jiao, H., & Zhao, H. (2017). Transcriptome sequencing and phylogenetic analysis of four species of luminescent beetles. *Scientific reports*, *7*(1), 1–11.
65. Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. springer.

66. Wigginton, J. E., Cutler, D. J., & Abecasis, G. R. (2005). A note on exact tests of Hardy-Weinberg equilibrium. *The American Journal of Human Genetics*, *76*(5), 887–893.
67. Wolf, J. B. (2013). Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Molecular ecology resources, 13*(4), 559-572.
68. Wright, B., Farquharson, K. A., McLennan, E. A., Belov, K., Hogg, C. J., & Grueber, C. E. (2019). From reference genomes to population genomics: comparing three reference-aligned reduced-representation sequencing pipelines in two wildlife species. *BMC genomics*, *20*(1), 1–10.
69. Wright, S. (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution*, 395–420.
70. Zhou, C., Liu, Y., Qiao, L., Liu, Y., Yang, N., Meng, Y., & Yue, B. (2020). The draft genome of the blood pheasant (Ithaginis cruentus): Phylogeny and high-altitude adaptation. *Ecology and evolution*.